Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Extracção Automática de Tópicos de Documentos

Por

Luís Filipe da Silva Teixeira, 29399

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para a obtenção do grau de Mestre em Engenharia Informática

Orientador: Prof. Doutor José Gabriel Pereira Lopes
Co – Orientador: Prof. Doutor Joaquim F. da Silva

Lisboa
2010

"As armas e os barões assinalados,
Que da ocidental praia Lusitana,
Por mares nunca de antes navegados,
Passaram ainda além da Taprobana,
Em perigos e guerras esforçados,
Mais do que prometia a força humana,
E entre gente remota edificaram
Novo Reino, que tanto sublimaram;

E também as memórias gloriosas
Daqueles Reis, que foram dilatando
A Fé, o Império, e as terras viciosas
De África e de Ásia andaram devastando;
E aqueles, que por obras valerosas
Se vão da lei da morte libertando;
Cantando espalharei por toda parte,
Se a tanto me ajudar o engenho e arte.
..."

Luís Vaz de Camões, Lusíadas, Canto I

Dedicatória

Aos meus Pais,

Maria Eduarda e Mario Teixeira

Agradecimentos

Não posso começar os agradecimentos, sem dar uma menção especial de agradecimento aos meus Pais, por tudo o que tem passado desde sempre e em especial nos últimos anos. À minha família que esteve sempre comigo e me acompanhou nestas etapas da minha vida.

Agradeço ao meu orientador, Prof. Doutor Gabriel Lopes, pela paciência que teve ao longo dos últimos meses. Paciência por todas as nossas "discussões" sobre o trabalho realizado nesta dissertação, as quais me possibilitaram atingir um novo nível de experiência e conhecimento. Agradeço-lhe ainda toda a força e motivação que me deu. Ao longo destes meses aprendi muito consigo. Além de orientador, considero-o acima de tudo um Grande Mentor e Amigo para o resto da vida. Foi um prazer realizar o trabalho desta tese na sua companhia.

Agradeço também ao meu Co-Orientador, Prof. Doutor Joaquim Ferreira da Silva, pela ajuda que me deu nos momentos em que precisei.

Um Agradecimento especial à Prof. Doutora Rita Ribeiro, coordenadora do CA3-UNINOVA, local onde cresci como profissional, investigador e acima de tudo como pessoa. Agradeço-lhe, ainda, ter-me ajudado a realizar um sonho de menino ao dar-me a hipótese de trabalhar na área do Espaço.

Agradecer ao Pessoal do CA3, que me acompanhou durante estes meses, e que aturou o meu "mau feitio" durante a realização da minha dissertação, quando as coisas não corriam pelo melhor à primeira.

Não posso deixar um agradecimento especial aos meus amigos de sempre, à minha namorada, que me apoiaram, e motivaram quando foi preciso.

A Todos o Meu Muito Obrigado.

Resumo

É amplamente conhecida a necessidade de se terem palavras-chave ou tópicos associados a documentos. Entende-se por palavras-chave ou por tópico (s) de um documento qualquer palavra ou multipalavra (uma sequência de 2 ou mais palavras) que, tendo um significado mais ou menos preciso, resume em si parte do conteúdo desse documento.
Neste trabalho pretendo desenvolver uma nova metodologia que aborda a problemática de extracção de palavras-chave. Para tal, trabalharei a extracção das palavras-chave trabalhando com palavras, multipalavras e prefixos de palavras com comprimentos predefinidos (5 caracteres). A utilização de prefixos permite trabalhar com línguas altamente flexionadas, servindo os prefixos tópico como sinalizadores de toda uma família de palavras e de multipalavras que poderão, nesse caso, ser promovidas a tópicos, sendo a extracção destes prefixos inovadora, relativamente ao estado da arte.

A extracção a realizar será baseada em estatística, o que possibilita trabalhar com textos de várias línguas, nomeadamente o Português, o Inglês e o Checo, que foram as línguas utilizadas neste trabalho. Pretendi melhorar os tempos de extracção de tópicos, recorrendo para isso à utilização de Suffix Arrays. Os resultados obtidos foram avaliados por pessoas externas.

É feita também uma comparação bastante exaustiva entre 24 métodos de extracção, alguns novos, propostos neste trabalho, outros propostos por outros autores.

Com esta dissertação pretendo fornecer uma nova ferramenta a trabalhos posteriores de sumarização de documentos, de Agrupamento ou indexação de documentos, de construção de ontologias.

Abstract

It’s widely known the need to have Keywords and topics associated to documents. A keyword or topic from a document is a word or multi-word (sequence of more than 2 words) that, having a more precise meaning, summarizes in itself part of the content of that document.
This work plan intends to develop a new methodology to work with the problem of automatically extracting key-words. For that, we intend to work this problem at the level of words, multi-words, and prefix of words with fixed length (4 and 5 characters). The use of word prefixes will allow us to deal with highly inflected languages, serving this kind of topic prefixes as a marker of an entire family of words or multi-words, which in that case, might be promoted to topics themselves, being the extraction of these prefixes innovative, relatively to the state of the art.
The extraction made is based on statistics, which will allow us to work with texts of several languages, namely Portuguese, English and possibly Czech that are the case study of this work.
We pretend to improve the extraction time of topics, and for doing that we made use of Suffix Arrays. The results were evaluated by external people.

It’s also made a very exhaustive comparison between 24 extraction methods, some new, proposed in this work, other proposed by other authors.

With this master thesis, we intend to offer a new tool, to posterior works that may be done in the areas of document summarization, clustering or Indexing of documents, and ontology construction.

Índice

Índice de Tabelas

Índice de Figuras

Glossário

Bag-of-Words⇔ Saco de Palavras

Bigrama⇔ Sequência de dois elementos de texto, normalmente palavras.

Cluster⇔ Grupo (Classe)

Clustering ⇔Agrupamento**,** método não supervisionado de identificação de grupos ou classes.

Corpus⇔ Colecção de textos provenientes de uma ou várias fontes distintas.

Corpora⇔ Múltiplas colecções de textos. Plural de corpus.

Formal Concept Analysis ⇔FCA⇔ Análise Formal de conceitos.

Information Retrieval ⇔ Recuperação de Informação

Lattice⇔ Malha

Lemmatization ⇔ Lematização

Links ⇔Ligações

Longest Common Prefix (LCP) ⇔ Prefixo comum mais longo

Multipalavra ⇔ Sequência de duas ou mais palavras, normalmente com significado e à qual se pode atribuir uma classe sintáctica.

Mutual Information ⇔ Informação Mútua.

N-grama de Palavras ⇔Sequência de n palavras.

Named Entities ⇔ Entidades com nome.

Noun Phrases ⇔ **S**intagmas nominais.

POS-Tagging, Part-of-Speech Tagging ⇔ Etiquetagem morfo-sintáctica.

POS-Tag, Part-of-Speech Tag ⇔ Etiqueta morfo-sintáctica.

Query ⇔ Pedido de informação, na área de recuperação de informação.

Stop Words ⇔ Palavras funcionais desprovidas de significado (artigos, preposições, …)

Script ⇔ Sequência de instruções a serem executadas sequencialmente.

String ⇔ **C**adeia de caracteres.

Unidades Lexicais Multipalavra, multipalavras, termos multipalavra ⇔ multiword units ⇔ Sequências de palavras que correspondem normalmente a nomes próprios, frases idiomáticas ou colocações com categoria gramatical.

Unigramas ⇔ Um elemento de texto, normalmente uma palavra.

Unipalavra ⇔ Uma palavra.

Tf-Idf, Term Frequency - Inverse Document Frequency ⇔ Frequência do termo - Inverso da frequência dos documentos onde o termo ocorre.

Trigrama ⇔ Sequência de três palavras ou mais elementos de texto, normalmente palavras.

Vector Space Model ⇔ Modelo Vectorial

Capítulo 1

Introdução

Entende-se por tópico ou palavra-chave de um documento qualquer palavra ou multipalavra (sequência de 2 ou mais palavras) que, tendo um significado mais ou menos preciso, resume em si parte do conteúdo desse documento de uma dada colecção.

São exemplos de tópicos altamente correlacionados os seguintes: agentes zoonóticos, zoonoses, zoonose, salmonela, organismo zoonótico, infecções zoonóticas, fiscalização sanitária, polícia sanitária, doenças zoonóticas, etc.

Outro exemplo pode ser verificado no ficheiro *pt_32006D644.html*[1] presente no corpus em português utilizado na realização deste trabalho, onde verificamos que um tópico altamente relevante é a palavras multilinguismo, que aparece também associado a "*domínio do multilinguismo*" e a *"peritos no domínio do multilinguismo"*, que são altamente correlacionados entre si e altamente descriminantes do conteúdo do documento em causa. Ver Figura 7.24

A extracção destes tópicos (ou palavras-chave) é útil numa variedade alargada de aplicações de que se destacam: a construção automática de ontologias, a sumarização de documentos, o agrupamento e a classificação de documentos, visando aceder mais facilmente e eficazmente à informação que realmente se procura.

[1] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006D0644:PT:HTML

Um dos requisitos das palavras-chave (ou tópicos) é o de deverem ser bons descritores do conteúdo dos documentos a que se referem. E, um bom descritor de um documento, é-o se for relevante no seu contexto pelo que importa definir o que é a relevância no âmbito em que se fala dele. Intuitivamente espera-se que a relevância de um termo (palavra ou multipalavra) num documento esteja relacionada com a sua frequência nesse documento e em documentos que tratem da mesma problemática, não devendo surgir, de preferência, em documentos que tratem de outros temas. Uma medida que modela este tipo de considerações (a relevância do termo) é a métrica Tf-Idf, que mede a frequência do termo (Tf, term frequency) multiplicada por um factor que tem em linha de conta o inverso da frequência dos documentos onde ocorre o termo (Idf, inverse of document frequency) na colecção considerada (ver secção 2.3.1.1).

No artigo de J.F. Da Silva e G. P. Lopes [1] são comparados os resultados da extracção de tópicos de documentos, para selecção de descritores de documentos, utilizando 3 medidas de selecção, onde se inclui também a métrica Tf-Idf, aplicadas a multipalavras previamente extraídas, pelo método *LocalMaxs*, apresentado no artigo [2]. No artigo [1] mostra-se que o Tf-Idf é uma medida que tem tendência a escolher termos multipalavra demasiadamente específicos[2]. Daí a necessidade de aqueles autores utilizarem duas outras métricas cujos resultados demonstraram a sua maior capacidade para a extracção de melhores descritores[3]. Contudo, este trabalho ficou-se pela extracção de descritores com mais de uma palavra, sendo-lhe impossível extrair descritores com uma única palavra, como seria o caso das "salmonelas", das "zoonoses" ou da "zoonose" acima referidas.

Por este motivo, o trabalho que foi levado a cabo, debruçou-se sobre a extracção de palavras-chave quer estas sejam palavras singulares ou multipalavras. Além disso, também estudei resultados que obtive tendo em consideração prefixos de palavras, com quatro e cinco caracteres, os quais também foram eleitos como relevantes, utilizando as mesmas métricas base que foram seleccionadas para eleger a relevância temática de palavras e de multipalavras.

Nos exemplos apresentados acima, "zoon" seria um prefixo de 4 caracteres que

[2]No trabalho, levado a cabo, contraria-se esta ideia, Tf Idf produz bons resultados, trabalhando tanto com palavras bem como com multipalavras. Talvez este resultado seja consequência dum filtro aplicado que só considera palavras com 6 ou mais caracteres.
[3]No trabalho desenvolvido, descrito nesta tese, prova-se também que as medidas apresentadas pelos autores, têm alguma dificuldade em diferenciar bons descritores.

ocorreria muito mais vezes do que qualquer das palavras isoladas (ou das multipalavras) que o contêm: "zonose", "zonoses", "zoonótico", "zoonótica", "zoonóticos" ou "zoonóticas" (ou , "agentes zoonóticos", "organismo zoonótico", "infecções zoonóticas" e "doenças zoonóticas). Em línguas altamente flexionadas como o Checo, em que os nomes podem chegar a ter 14 formas diferentes (7 singulares e 7 plurais, uma para cada um dos casos) e os adjectivos podem chegar a ter 42 formas diferentes (3*7 singulares e 3*7 plurais, uma para cada um dos três géneros possíveis, masculino, feminino e neutro), pensámos e comprovámos, no trabalho realizado, que uma abordagem com base em prefixos de palavras poderia altamente produtiva. Se, pretendêssemos estender a metodologia a línguas orientais, como o Chinês ou o Japonês, trabalharíamos provavelmente com sequências de 2 caracteres, eventualmente 3, ou mesmo um único carácter porque, nestas línguas, não existe o espaço em branco como separador de palavras e porque há palavras de conteúdo que se escrevem com um único carácter.

Tal como ficou escrito acima, neste trabalho extraímos palavras de comprimento mínimo de 6 caracteres, sendo este parâmetro configurável consoante o que quisermos avaliar (ver capítulo 3), multipalavras (previamente identificadas utilizando a metodologia referida em [2] por Silva et al) e prefixos de palavras com 5 caracteres que possam ser considerados como relevantes para o tópico em discussão nos documentos onde existirem. No que se refere às métricas a utilizar para detectar as unidades textuais relevantes (palavras, multipalavras e prefixos), foram utilizados o Tf-Idf, já mencionado, um adaptação da métrica Rvar utilizada em [1], o Chi-quadrado[4] [3] [4], o Phi-quadrado e a Informação Mútua [5]. Diversas variantes foram desenhadas para melhor comparar, em condições de igualdade, as vantagens e desvantagens de cada uma das métricas.

Ao fazer-se isto, constatou-se que algumas medidas nos davam resultados que não permitiam uma clara identificação de um bom descritor, visto que atribuíam o mesmo valor às trinta ou quarenta primeiras palavras ou multipalavras. Isto acontece claramente com o Rvar e com a Informação Mútua, bem como com algumas das variantes destas medidas. Outra das conclusões, foi a de que o Tf-Idf o Phi-Quadrado, juntamente com algumas das suas variantes, são as medidas que produzem resultados

[4] O Chi quadrado é semelhante ao Phi- quadrado e para efeitos de avaliação dá os mesmos resultados que o Phi-quadrado.

mais interessantes. Mais informação será encontrada no Capítulo 3.

1.1 Motivação

Ao pretender extrair também as palavras que caracterizam o conteúdo de qualquer documento, pretendi estender o trabalho realizado por J. F. Silva e Lopes [1] a este tipo de unidade textual e comparar os resultados obtidos em [1], com os que obtive ao longo deste trabalho. Vi esta necessidade porque algumas vezes uma boa palavra pode ser um descritor altamente objectivo do conteúdo concreto de um documento, como já mencionado no início da Introdução, "*multilinguismo*" é uma palavra, mas denota o conteúdo de um dos documentos estudados de uma forma inequívoca.

Uma outra ideia, que contribui para a elaboração desta dissertação, e também estendendo o trabalho [1], foi o de usar prefixos de palavras, como possíveis descritores de documentos, e a relação destes com as palavras.

Veja-se o caso do prefixo *"multi"*, que ao ser prefixo de multilinguismo, também é prefixo de *"multilinguista", "multilinguistas", "multilculturais".* O que nos deu a ideia de propagar o valor da medida de importância do prefixo atribuindo-o às palavras que fossem iniciadas por esse prefixo. A este processo, foi dado o nome de *"Bubbling"*, como se fizéssemos "borbulhar" os valores das medidas de relevância dos prefixos para as palavras que os contêm.

Outra situação, que motivou a realização deste trabalho foi a de como estender a ideia de "*Least"*, que em [1] é aplicado somente a multipalavras. Este processo fez com que admitíssemos e assumíssemos que o *"Least"*, que em [1] media o valor mínimo de uma medida (Rvar) das palavras extremas (direita e esquerda) de uma multipalavra, passasse a medir o valor desse medida para a própria palavra. Ou seja, a palavra passou a ser tomada como uma multipalavra cujos extremos são iguais à própria palavra.

Ao utilizar as medidas *Tf-Idf*, *RVar*, φ^2 e *IM*. (ver secções de 2.3.1.1 a 2.3.1.5) para identificar o grau de importância relativamente a cada documento, aplicadas não só a palavras, mas também a prefixos e multipalavras, e ao utilizar também variantes destas medidas, resultantes de conjugações de formas diferentes de fazer sobressair termos

relevantes, nomeadamente fazendo uso da técnica de *Bubbling* (descrita na secção 3.2.2), fazendo uso da Mediana do comprimento das palavras e das palavras constituintes de multipalavras (ver secção 3.2.4 e 3.2.5) para dar maior diferenciação a esses termos, usando a ideia de aplicar o operado *Least* já referido (ver secção 3.2.1), a todas as métricas base e aplicando uma combinação entre a métrica base, o operador *Least* e a técnica de *Bubbling* (ver secção 3.2), estabeleci assim um campo de experimentação vasto para poder comparar extensivamente todas estas medidas e respectivas variantes no processo de extracção de termos chave, visando a avaliação final dos resultados obtidos. As métricas *Tf-Idft* e *RVar e a Informação Mútua* já foram utilizadas neste tipo de experiência, como veremos em algumas subsecções do capítulo 2. Mas o φ^2 (e o χ^2 com resultados equivalentes) são medidas muito úteis e muito utilizadas para a selecção de *features* mais relevantes para serem utilizadas por classificadores de texto [5] [6], nunca foram, tanto quanto sei, utilizadas neste tipo de experiência. Todas as variantes de medidas que foram criadas e aplicadas neste trabalho, nunca foram aplicadas em nenhum contexto anteriormente a este trabalho.

Como consequência do ponto anterior, teve-se de pensar numa maneira de possibilitar a comparação em simultâneo de palavras e multipalavras, visto que faria pouco sentido fazer uma avaliação somente para palavras, e outra somente para multipalavras, assim decidiu-se fazer a junção numa só estrutura das palavras e multipalavras e fazer a avaliação e a extracção dos termos mais relevantes desta estrutura.

Com estes resultados, foi-me possível fazer a comparação entre os resultados que obtive com estas métricas na extracção de prefixos, de palavras e de multipalavras relevantes na identificação dos tópicos dos documentos onde ocorrem. Como consequência foi possível observar que, muitas vezes, os melhores descritores são palavras singulares.

Desta forma, foi também possível observar que, uma escolha arbitrária como é feita em [7], onde os autores optaram por avaliar as dez multipalavras mais relevantes e as três palavras mais bem cotadas, não é a forma mais adequada para tratar este problema. De facto, há documentos em que são palavras maioritariamente que descrevem os conteúdos dos documentos, e como consequência não é adequado fixar, à partida um número de palavras e outro de multipalavras para descrever o documento.

Os resultados apresentados em [8] com recurso à utilização de Suffix Arrays, motivaram a escolha desta estrutura de dados para utilizar neste trabalho, acreditando poder provar a sua grande utilidade e eficácia, como explico melhor na secção 2.9 especificamente dedicado às Suffix Arrays. De facto, ao recorrer a elas terei a capacidade para determinar quase instantaneamente as frequências de prefixos, de palavras e de multipalavras distribuídas por cada um dos documentos onde ocorrem.

1.2 Solução Desenhada

No âmbito do trabalho que desenvolvi e que culminou a escrita desta dissertação, pretendi como já referido ao longo da Introdução (secção 1), extrair automaticamente termos-chave (ou tópicos) de documentos, que sejam bons descritores do conteúdo desses mesmos documentos. Além da extracção de palavras e multipalavras descritoras, num exemplo como o apresentado na secção 1, e como verificado nos resultados obtidos, "*multi*" seria um prefixo de 5 caracteres altamente discriminante do documento em causa. Com este tipo de informação extraída, tornou-se possível procurar palavras e multipalavras que contenham os prefixos seleccionados. No exemplo citado, seria o caso de "multilinguismo", "multilinguista" ou "multiculturais" entre outros termos. Ao seguir esta linha de trabalho, aumentámos a cobertura sobre as palavras e multipalavras que podem ser representativas do documento e que eventualmente podem não ser extraídas quando o método é aplicado exclusivamente a palavras ou a multipalavras e não dispomos de dicionários para reduzir esses termos chave à sua forma singular ou plural, consoante se considere que o singular ou o plural é mais representativo do assunto. Esta opção mostrou-se adequada numa língua altamente flexionada como é o caso do checo. Com esta opção diminuímos a precisão mas aumentamos a cobertura, mesmo para Português e para Inglês.

Assim, parte do trabalho foi destinado a extrair listas de palavras e multipalavras e prefixos ordenados por grau de importância. Depois foi feita a conjugação entre as várias listas, de forma a no final, para todas as métricas e variantes, poder escolher, para o caso das palavras e das multipalavras, as 25 melhores, por documento, para serem avaliadas em 5 documentos escolhidos aleatoriamente, para que se obtenha uma análise crítica sobre os resultados obtidos na extracção automática efectuada.

No caso dos prefixos, estes foram extraídos e a sua importância repercutiu-se sobre as palavras e multipalavras que os continham. A esta técnica chamei de *bubbling*. O que possibilitou fazer o cálculo de outras variantes de medidas (ver capítulo 3 secção 3.2). Depois dos termos extraídos e avaliados foi feita uma extracção de valores de precisão para os 5, 10, 15 e 20 melhores. Tendo estes resultados, foi feita uma avaliação no grau de concordância entre pares de avaliadores recorrendo à estatística kappa (ver secção 2.8.3)

Para alcançar esta potencialidade de extrair palavras, multipalavras e prefixos relevantes, e no caso dos prefixos, extrair palavras ou multipalavras que contenham esses prefixos, recorri ao uso de *Suffix Arrays[8]*, por esta estrutura permitir trabalhar com todas as variações já faladas até aqui, palavras, multipalavras e prefixos de um documento ou de uma colecção, permitindo, em particular, determinar eficientemente a frequência dessas unidades lexicais na colecção e em cada um dos seus documentos.

1.3 Principais Contribuições

Uma das principais contribuições deste trabalho foi propor novas métricas para a extracção de palavras e multipalavras chave descritoras do conteúdo de documentos de uma dada colecção. Além disso comparo os resultados das 24^5 métricas de extracção de termos chave.

Os resultados foram avaliados por pares de avaliadores independentes (consultar capítulo 3 sobre contribuições e trabalho realizado). Utilizei estatística Kappa (ver secção 2.8.3) para medir o grau de concordância entre as avaliações atribuídas por cada um desses avaliadores e medir o grau de credibilidade que as avaliações feitas têm.

Nos trabalhos estudados, a identificação de prefixos de 5 caracteres que sejam tematicamente importantes não é feita. Sendo que esta também é uma das contribuições deste trabalho e a avaliação feita em duas línguas morfologicamente ricas como o Português e o Checo (ver os resultados obtidos no capítulo 4), comprovou-se que a sua aplicação traz resultados interessantes. O Inglês apesar de ser morfologicamente pobre em comparação com as duas línguas nomeadas anteriormente também beneficiou com o uso desta alternativa.

[5] 30 se considerarmos o Chi-Square.

Não experimentei o uso de sequência de caracteres mas antevejo também a possibilidade do que se pode passar com línguas asiáticas como o Chinês ou Japonês onde utilizaria no máximo sequências de dois ou três caracteres, ou do alemão onde se poderão utilizar cadeias de 4 ou 5 caracteres, não necessariamente prefixos Crê-se que, desta forma, para as línguas indo-europeias, é possível aumentar a cobertura dos resultados obtidos sem diminuir o grau de precisão que já se obtêm [1] De facto, ao que permitir capturar as palavras ou multipalavras que, de outra forma, poderiam facilmente não ser apanhadas devido a frequências de ocorrência muito baixas, se só levasse em linha de conta a utilização de palavras ou multipalavras ocorrendo de facto nos documentos, com este trabalho contribui-se para aumentar a cobertura, sem ter diminuído a precisão.

Ao comparar explicitamente várias métricas que foram utilizadas na selecção das palavras-chave a extrair, e não tendo havido anteriormente nenhuma comparação entre estas métricas para os efeitos pretendidos neste trabalho, contribuí assim para um conhecimento mais profundo sobre este assunto podendo daí inferir qual(ais) o(s) método(s) melhor(es) a utilizar e as situações mais adequadas para o fazer.

Outra das contribuições, será o de abordar este problema utilizando uma estrutura de dados adequada para o fazer, as Suffix Array (ver secção2.9), que acarretou maior velocidade no processo de extracção de termos chave. Convém dizer que a sua utilização não é prática corrente em nenhum dos trabalhos estudados e apresentados no Estado da Arte (capítulo 2).

1.4 Organização da Dissertação

Esta dissertação está dividida da seguinte forma: no capítulo 2 serão apresentados diversos trabalhos, relacionados com o tema desta dissertação, que constituem actualmente o “estado da arte” na extracção multipalavras, e nas possíveis aplicações que fazem uso de termos relevantes no sentido de descrição dos tópicos do texto em análise. Este capítulo está dividido em várias secções, onde descrevo temas como Representação de Documentos, onde apresento várias formas de como um documento pode ser representado computacionalmente. Uma outra secção trata Descritores de Documentos, e como esta definição de descrição deve ser diferenciada de sumarização.

Uma terceira secção trata de Metodologias de Extracção, na vertente Estatística, onde a extracção de termos de um documento é efectuado que tem por base análises estatísticas de documentos. Nesta secção trata-se ainda vertente não estatística, onde a extracção é efectuada recorrendo a outros mecanismos como a etiquetagem morfossintácticas.

Seguem-se duas secções, uma sobre Extracção de Palavras e outra sobre Extracção de Multipalavras. Em cada tema apresento alguns trabalhos realizados no âmbito desses temas, ou que nalguma componente se relacionam com o tema charneira desta dissertação.

Após esta secção apresentam-se áreas de aplicação das metodologias apresentadas.

Nas secções finais do capítulo 2 apresentam-se Medidas de avaliação de resultados, algumas notas finais sobre o capítulo, e a estrutura de dados utilizada neste trabalho.

No capítulo 3 são apresentadas as contribuições desta dissertação, onde se apresentam mais em detalhe algumas das variantes das métricas base, sobre as quais se podem fazer análises interessantes. No capítulo 4 serão apresentados e comentados os resultados obtidos pelas várias métricas e as suas variantes, comparando-os com os resultados obtidos com a implementação dos outros métodos analisados neste trabalho. Finalmente, no capítulo 5 serão apresentadas as conclusões e o trabalho futuro.

Capítulo 2

Estado da arte

Nos últimos anos houve um aumento de importância e de necessidade de análise e compreensão automática do conteúdo de textos dado o crescimento enorme da informação em suporte digital e da necessidade de se ter acesso fácil à informação neles contida considerada necessária e adequada.

Este factor levou ao aumento da utilização de diversas ferramentas e metodologias desenvolvidas para ajudar na resolução do problema de processamento de documentos de texto visando diversas aplicações de que destaco a classificação automática, o reconhecimento de entidades com nome (*named entities*), a sumarização de documentos, o agrupamento de documentos, a indexação de documentos, e a recuperação de informação

Neste capítulo referencio vários trabalhos, de forma faseada, de diversos autores realizados nas áreas de aplicação já mencionadas no capítulo 1 da Introdução, designadamente: representação de documentos, descritores de documentos, entre outros. Trabalhos que no seu conteúdo fazem uso da extracção e da identificação de termos com importância, sendo esta parte, sempre uma componente de processos mais complexos.

Abordarei também, metodologias de extracção de termos chave, palavras e multipalavras.

Depois, e por uma questão de completude, apresentarei a seguir métricas de avaliação dos resultados. Há também a apresentação das Suffix Arrays que é a estrutura de dados eleita para utilização neste trabalho.

2.1 Representação de Documentos

A representação dos documentos poderá ser realizada de várias formas. Há uma forma de representação mais usual em trabalhos da natureza deste e que é a de o documento ser representado por um vector em que os constituintes são as palavras que constituem o documento. Esta é a representação saco de palavras.

Os documentos podem também ser representados pelas multipalavras lá contidas que os constituem. Entende-se por uma multipalavra uma sequência não interrompida de palavras que se deseja que tenham necessariamente um significado, como seria o caso de "câmara escura", "máquina fotográfica", "indústria cinematográfica".
Estas multipalavras ou são extraídas tendo em linha de conta informação morfo-sintática de cada um dos seus constituintes das frases do documento, não sendo por isso a sua extracção independente da língua [9, 10], ou são extraídas tendo em linha de conta o grau de coesão estatística entre as palavras constituintes dos documentos em análise [2], sendo neste caso a sua extracção independente da língua.
Por exemplo, no seguinte texto:

> "A Câmara Municipal de Murça organiza o segundo Raid de Fotografia Digital."

Podemos encontrar em sequências de 2-gramas de palavras, o seguinte:

> A Câmara; Câmara Municipal; Municipal de; de Murça; Murça organiza; organiza o; o segundo; segundo Raid; Raid de; de Fotografia e Fotografia Digital.

Do mesmo exemplo, podemos encontrar as seguintes multipalavras:

> Câmara Municipal; Câmara Municipal de Murça; Raid de Fotografia Digital e Fotografia Digital.

Na realização deste trabalho, os documentos são representados por palavras, por multipalavras e por prefixos de palavras, sendo que os prefixos não são directamente apresentados aos avaliadores. São antes utilizados para propagarem as medidas da sua importância às palavras e multipalavras que os contêm via uma técnica que designamos por *Bubbling*. São, deste modo, utilizados internamente para realização de cálculos, cujos pormenores podem ser vistos no capítulo 3.

Trabalhos existem, onde os documentos são representados por Web-Snippets [11-13], ou por parágrafos [14]. Sendo qualquer destas representações reduzidas depois à consideração das palavras lá existentes e também das multipalavras constituintes [12].

2.2 Descritores de Documentos

Um descritor de um documento é um termo que capta a essência do conteúdo de um documento. Importa desde já fazer uma distinção clara, entre o que entendo por descritores de documentos e por sumarização (ver secção 2.6.2) de documentos na medida em que, em algumas circunstâncias, pode haver confusão entre o que é uma coisa e o que é outra.

A Sumarização de Documentos, é o processo de criação de uma versão mais curta de um texto, sendo que esta versão mais curta, contém os pontos relevantes do texto original. Nalguns casos essa versão mais curta é um parágrafo ou uma frase retirada(o) do documento a sumarizar. Mas noutras aplicações pode reduzir-se a sumarização à extracção de termos chave.

Quando falamos de descritores de documentos, estamos a falar de palavras-chave ou de termos chave, que por si só dão uma clara ideia do conteúdo de um documento, e é esta a ideia base do trabalho desenvolvido nesta dissertação.
Para a realização desta dissertação, tomei como ponto de partida o trabalho desenvolvido por Joaquim F. da Silva et.al. no trabalho [1], onde se aborda o tema de descritores multipalavra de documentos, como já referido anteriormente. Em [1], são utilizadas expressões multipalavras extraídas, recorrendo ao algoritmo LocalMaxs [2], em conjugação com a medida estatística SCP e com a normalização do SCP através da aplicação do FDPN (Fair Dispersion Point Nomalization) [15]. Podemos ver mais informação sobre o SCP, o FDPN e o algoritmo LocalMaxs, na secção 2.5 sobre a extracção de multipalavras. Após a extracção das expressões relevantes, são aplicadas medidas estatísticas, *Tf-Idf*, *RVar*, *LeastRVar* e *LeastRVarLen* (ver secções de 2.3.1.1 e 2.3.1.2), para se ordenarem por ordem de importância, de acordo com a medida utilizada, as expressões multipalavra obtidas, assumindo que as mais bem classificadas poderão ser consideradas descritores de documentos.

Assim, um dos meus objectivos foi o de estender este trabalho, como já referido anteriormente no capítulo 1, trabalhando também com palavras e com prefixos, que não haviam sido abordadas em [1].

No trabalho que levei a cabo utilizei para a extracção 4 medidas base, o *Tf-Idf*, a medida *Rvar*, o φ2 e acrescentei ainda a Informação Mútua. Para mais pormenores sobre estas medidas ver secção 2.3.1.

Estas medidas foram aplicadas na extracção das palavras, das multipalavras e dos prefixos mais descritores do conteúdo dos documentos e na sua análise. Mas sentiu-se a necessidade de poder fazer uma comparação com a variante *LeastRvar* apresentada em [1]. Dessa necessidade surgiu o desenvolvimento de cinco variantes, uma para cada medida base. Estas variantes possibilitaram a criação da versão *Least* para cada medida (ver secção 3.2.1). As outras variantes surgiram de outras necessidades:

- Como a de ter em consideração a mediana do comprimento das palavras constituintes de uma multipalavra (ver secções 3.2.4 e 3.2.5) já que em [1] o comprimento médio das palavras constituintes de multipalavras foi uma característica testada para a extracção. Mas no trabalho que desenvolvemos, como tratamos palavras e multipalavras em simultâneo, a escolha do uso da mediana recai sobre a análise do trabalho [7] onde o uso da mediana foi também testado.

- Outra situação, advém de como poderíamos relacionar os prefixos e as palavras, o que levou ao *"Bubbling"* (ver secção 0), processo de atribuir a uma palavra, o valor da medida tida pelo prefixo da palavra.

Em resumo, o trabalho [1] despoletou a necessidade de comparar exaustivamente várias métricas (*Tf-Idf, Phi-Square, Rvar e Informação Mútua*), várias representações dos documentos (palavras, multipalavras e prefixos) e medir a precisão e a cobertura atingidos por cada uma dessas métricas e das variantes criadas.

Um outro trabalho [16] tem como um dos componentes, a extracção de descritores de documentos, mas sendo que aqui não são utilizadas multipalavras como no trabalho

anterior, mas sim noun-phrases[6]. Neste trabalho os autores propõem uma técnica para seleccionar automaticamente sintagmas nominais (*noun phrases*) como descritores de documentos para conseguirem construir uma "*FCA – Based IR Framework*", onde FCA [17] significa Análise Formal de Conceitos (Formal Concept Analysis) e "*IR Framework*" sugere que o trabalho é feito no âmbito da recuperação de informação.

A proposta que os autores apresentam é composto por cinco passos,

- O texto dos documentos e dos pedidos de informação são indexados e comparados num *"Vector Space Model"* utilizando para isso os pesos dados pela medida *Tf-Idf*. Para um dado pedido de informação, uma lista ordenada de documentos é criada a partir deste modelo.
- Os primeiros n documentos nesta lista são examinados para extrair dos termos pertencentes aos documentos, um conjunto de k descritores óptimos de acordo com uma determinada medida de peso.
- *"Formal Concept Analysis"* é aplicada ao conjunto de documentos como sendo objectos formais, onde os atributos formais de cada documento são um subconjunto dos k descritores que são contidos no texto.
- Além da caracterização intencional de cada nó conceito, uma descrição adicional é construída com sintagmas nominais mais salientes que incluam um ou mais termos do pedido de informação. Esta caracterização é usada para aumentar a descrição dos nós na malha conceptual utilizada no sistema dos autores.
- A malha anotada resultante á apresentada ao utilizador que pode navegar os primeiros n resultados atravessando a malha podendo depois refinar o pedido de informação a qualquer momento.

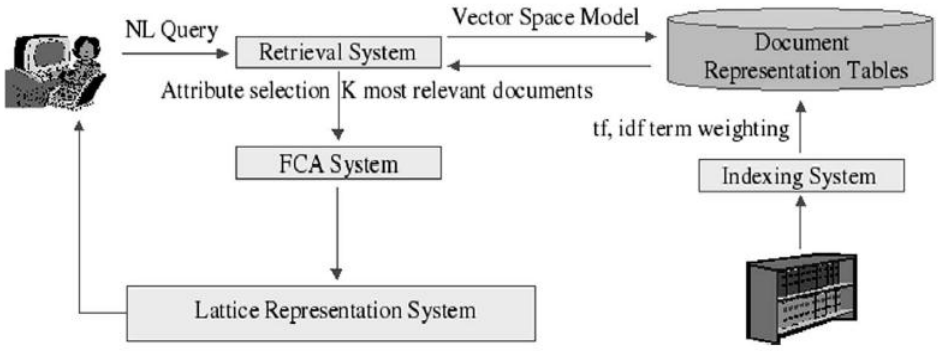


Figura 2.1 - Ilustração do Sistema proposto pelos autores no trabalho [16].

[6] Sintagmas nominais.

Por forma de completude, descreve-se Análise Formal de Conceitos como sendo um método particular de análise de dados e de representação de conhecimento [18] [19] que se baseia numa malha conceptual[7]. A ideia base no FCA é a de que é possível argumentar que uma malha conceptual é uma ferramenta eficiente para várias aplicações nomeadamente o agrupamento de conceitos, vertente que é trabalhada nos trabalhos [16] e [17], onde os autores argumentam ainda que outras vantagens de utilizar uma FCA em vez dos tradicionais algoritmos de Clustering de documentos é a de a FCA fornecer uma descrição de cada classe de documentos que pode ser utilizada para refinamento ou modificação, tornando assim as classes mais interpretáveis. E como os resultados vêm organizados numa malha em vez de aparecerem hierarquicamente organizados, e sendo esta a organização mais natural quando múltiplas classificações são possíveis, estes factos facilitam a possibilidade de se recuperar de más decisões enquanto se navega nessa malha para encontrar informação relevante. Tomemos como exemplo de uma FCA o que nos é apresentado em [20]. Primeiro observemos uma imagem de um pequeno contexto formal. Os elementos à esquerda são objectos enquanto os elementos no topo da tabela são atributos ou propriedades desses objectos.

	president (pr)	prime-minister (pm)	european union (eu)	kingdom (k)	islamic rules (ir)
Belgium (B)		X	X	X	
Portugal (P)	X	X	X		
Pakistan (PK)	X	X			X
Iran (I)	X				X
Saudi Arabia (A)				X	X

Figura 2.2 – Um conceito formal de "estados"

Exemplo retirado de [20]

Podemos construir uma malha formal de conceitos que consiste em duas dimensões linguísticas:

- Uma dimensão é a definição de intenção[8], ou seja, um conjunto de contextos léxico sintácticos similares com as mesmas restrições de selecção.

[7] Também conhecida como *Galois Lattice*
[8] Intension definition

- A outra é a extensão, que é o conjunto de palavras que aparece nos contextos e que satisfaz os requisitos semânticos.

Assim, da tabela da Figura 2.2, é possível retirar os conceitos formais e correspondente informação e construir a seguinte malha.

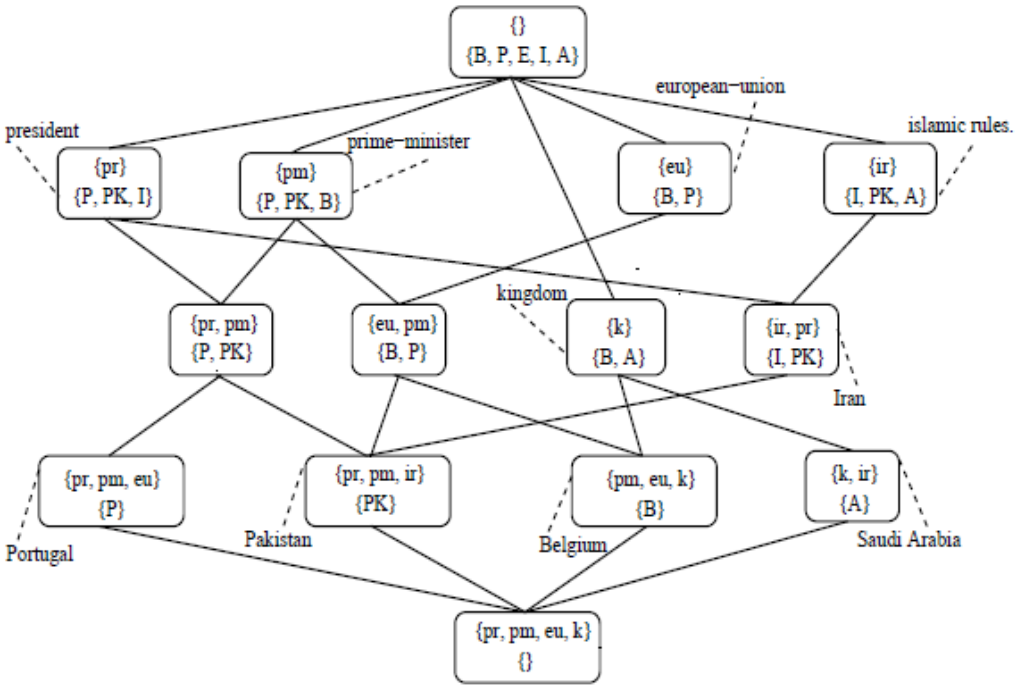


Figura 2.3 – Malha formal de conceitos do contexto formal identificado na Figura 2.2

Exemplo retirado de [20]

Para avaliarem os seus resultados, os autores de [18] apresentam três estratégias diferentes para seleccionar elementos frásicos (sintagmas nominais), que posteriormente são avaliados. Para avaliar os resultados que obtêm os autores definem as seguintes medidas, "*Minimal Browsing Area*" [17], que é a parte mínima da malha de conceitos que um utilizador deve consultar a partir do nó raiz até chegar aos conceitos relevantes, minimizando o número de documentos irrelevantes que tem de ser inspeccionados para obter toda a informação relevante. Recorrem também ao uso da "*Lattice Distillation Factor*" [17], sendo que esta é definida como sendo o ganho potencial de precisão entre a malhar e a lista de ordenada de conceitos, e é definida como

$$LDF(C) = \frac{Precision_{MBA} - Precision_{RL}}{Precision_{RL}} \quad , \qquad (2.3)$$

Onde C, é um conjunto de nós da malha conceptual, onde documentos estão marcados como sendo relevantes ou não relevantes para uma dada *query*. $Precision_{RL}$ é a precisão da *"Ranked List" e* $Precision_{MBA}$ é e a precisão da *"Minimal Browsing Area*". Segundo os autores a "*Minimal Browsing Area*" e "*Lattice Distillation Factor*" podem ser ambas aplicadas a agrupamentos hierárquicos ou qualquer outro agrupamento de resultados. A única dificuldade que os autores apontam ao de calcular a "*Lattice Distillation Factor*" é a de encontrar a "*Minimal Browsing Area*" para uma determinada malha. Para ultrapassar esta dificuldade criaram um grafo associado onde todos os nós são conceitos relevantes, e onde o custo associado a cada arco está relacionado ao número de documentos irrelevantes que serão acedidos atravessando esse arco. Seguidamente calculam uma "minimal span tree" para este grafo, que lhes dará a "*Minimal Browsing Area*". Podemos ver um exemplo do cálculo de uma LDF na seguinte Figura 2.4.

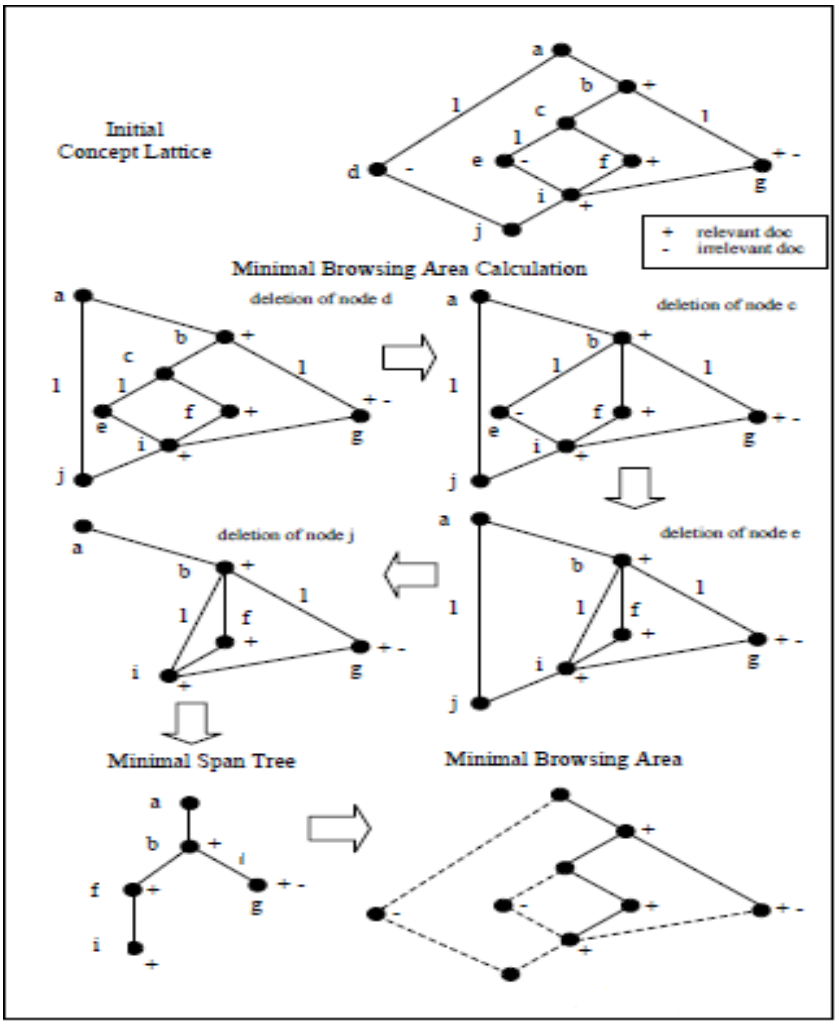


Figura 2.4 – Cálculo de uma "Lattice Destallation Factor"

Exemplo retirado de [17]

No exemplo da Figura 2.4, a Precision(ranked list) = 4 / 7 , a Precision de (Minimal Browsing área) = 4 / 5. Com estes valores é possível calcular o

$$LDF = (4/5 – 4 / 7) / (4/7) * 100 \% = 40 \%.$$

Estes autores utilizaram esta metodologia, para alcançarem um sistema que combina um motor de pesquisa de texto livre, como o Google, com uma malha conceptual para organizar os resultados de uma *query*.

Há que salientar ainda que este trabalho recorre a algumas ferramentas dependentes da língua, nomeadamente eliminação de palavras sem significado semântico[9], lematização[10], etiquetagem morfo-sintática e reconhecimento de padrões sintácticos, para extrair multipalavras (normalmente sintagmas nominais), com mais precisão e cobertura mas também requerendo conhecimento da língua. Sobre alguns destes temos voltarei a entrar em algum pormenor na secção 2.3.2.

2.3 Metodologias de Extracção

As abordagens são divididas em dois grupos: as que utilizam métodos estatísticos e as que utilizam outras abordagens essencialmente não estatísticas. Existe na literatura consultada diversas métricas para calcular o peso das palavras extraídas, apresentam-se algumas delas nas próximas secções, dando especial ênfase às estatísticas porque estamos interessados em métodos independentes de línguas.

2.3.1 Estatísticas

Quando se fala em abordagens estatísticas, estas podem basear-se numa abordagem que define um termo como uma palavra simples, e sabe-se que as palavras podem ser pré-processadas, o que pode incluir, entre outras operações, a de excluir palavras que não são relevantes em termos de extracção de informação, nomeadamente artigos, preposições, conjunções, entre outras palavras sem significado semântico relevante. Visto serem estas as mais frequentes e que ocupam cerca de quarenta por cento das ocorrências, mesmo sabendo que por vezes nestes quarenta por cento podem ser

[9] Stop Words
[10] É o processo de agrupar as diferentes formas flexionadas duma palavra resumindo-as a uma forma básica, para que possam ser analisadas como um único elemento.

incluídos alguns termos que contenham algum significado. Mas as abordagens estatísticas baseiam-se sobretudo em medidas de frequência e outras mais específicas que apresento nas subsecções seguintes.

2.3.1.1 Tf Idf

O Tf-Idf (*Term Frequency - Inverse Document Frequency*) foi inicialmente apresentado em [21] por Salton e Buckley.

Trata-se de uma métrica de cálculo de relevância de termos bastante utilizada nas áreas de Recuperação de informação (Information Retrieval), de Extracção de Informação e de *text-mining*. Permite medir o quão importante um termo (palavra, multipalavra ou prefixo) é num determinado documento em relação a outros termos ocorrendo nesse e noutros documentos da colecção ou corpus considerado para estudo. Esta métrica é obtida pela multiplicação de duas partes distintas, *Tf* e o *Idf*.

A primeira componente, *Tf*, mede o número de vezes que um termo (uma palavra, uma multipalavra ou um prefixo, ou qualquer outra sequência de caracteres) ocorre num determinado documento, ou seja, representa a frequência do termo. Esta contagem é depois normalizada para prevenir que as palavras em documentos muito extensos obtenham valores de *Tf* muito elevados e, em consequência, pouco rigorosos em relação a outros documentos mais reduzidos. A equação 2.1 mede, portanto, a probabilidade de um termo i ocorrer num documento j.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \quad (2.4)$$

onde $n_{i,j}$ é o número de vezes que o termo i ocorre no documento j; o denominador desta equação denota o somatório da frequência de todos os termos do documento, isto é, por outras palavras, o tamanho do documento j.

A componente *Idf* mede a importância geral de um determinado termo *ti* numa colecção de documentos. É definida com base na contagem do número de documentos em que esse determinado termo ocorre, como se pode ver na equação (2.5).

$$idf_i = \log \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|}, \quad (2.5)$$

Onde $|D|$ representa o número total de documentos no corpus ou colecção, e $\left|\{d_j : t_i \in d_j\}\right|$ o número de documentos onde o termo ti ocorre pelo menos uma vez, isto

é, $n_{i,j} \neq 0$. Nesta componente há que ter em linha de conta que é insensível à distribuição das ocorrências pelos diferentes documentos e valoriza excessivamente as ocorridas por lapsos ortográficos e todas as ocorrências raras, em particular se a colecção de documentos for muito grande.

Se olharmos com algum cuidado para a equação (2.6), que define *Tf-Idf* , constata-se que ocorrências únicas leva a baixos valores resultantes da equação (2.4), em particular se os documentos onde aparecem forem grandes, e a um máximo no valor resultante de (2.5), especialmente no caso de colecções muito grandes.

$$Tf - Idf = tf_{i,f} \,.\, idf_i \ , \quad (2.6)$$

Com base nesta medida, torna-se possível comparar entre documentos diferentes a importância obtida para cada termo, em particular se as colecções de documentos com que se trabalha não forem muito grandes ou, pelo menos, se os tamanhos dos documentos constituintes não forem demasiado pequenos.

No trabalho realizado, esta medida foi utilizada quando a representação dos documentos é feita com base em palavras ou prefixos ou multipalavras. Com as respectivas adaptações, mais propriamente na componente $tf_{i,j}$.

Quando a representação é feita por multipalavras, $n_{i,j}$ da equação (2.4), representa o número de vezes que a multipalavra i ocorre no documento j. Quando a representação é feita com base em prefixos, $n_{i,j}$ da equação (2.4), representa o número de vezes que o prefixo i ocorre no documento j. De forma análoga são tratadas as palavras. O denominador desta componente para os casos das palavras e dos prefixos é o somatório da frequência de todos os termos do documento, isto é, por outras palavras, o tamanho do documento j. No caso das multipalavras, para sermos mais correctos deveríamos ter diminuído àquele denominador, o número de palavras de cada multipalavra menos um. Isto justifica-se porque o número de possíveis multipalavras constituídas por N palavras existentes num documento é igual ao número de palavras desse documento menos (N-1). Contudo porque os documentos com que trabalhámos eram todos de tamanho superior a setecentas palavras, optámos por não complicar mais os cálculos e por não alterar aquele denominador. Convém acrescentar que trabalhámos com multipalavras de cinco palavras no máximo.

Na componente idf_i não existe nenhuma adaptação e a componente é calculada de forma idêntica para todas as variantes de representação de documentos adoptada.

Na experimentação realizada, ver Capítulos 3 e 4, pudemos verificar que a métrica *RVar* (secção 2.3.1.2) e *Informação Mútua* (secção 2.3.1.5) são também muito sensíveis aos lapsos ortográficos ou a ocorrências raras.

2.3.1.2 Rvar , LeastRvar e LeastRvarLen

De acordo com o trabalho realizado em [1], onde só se avaliaram multipalavras, os autores afirmam que a métrica Tf-Idf não privilegia necessariamente as expressões relevantes multipalavra mais fortes[11]. Assim, para colmatar esta aparente fraqueza da medida *Tf-Idf*, em [1] propuseram uma nova métrica *LeastRVar(.)*. Vocacionada para promover ou despromover multipalavras extraídas automaticamente sem recurso a qualquer conhecimento linguístico [2].

$$LeastRVar(RE_i) = least\left(RVar\left(lmostw(RE_i)\right), Rvar\left(rmostw(RE_i)\right)\right), \quad (2.7)$$

onde

$$Rvar(W) = \frac{1}{\|D\|} \sum_{d_i \in D} \left(\frac{p(W,d_i) - p(W,.)}{p(W,.)}\right)^2, \quad (2.8)$$

e onde $p(W,.)$ tem o significado de probabilidade média da palavra W tendo em conta todos os documentos e $Rvar(.)$ é aplicado à palavra mais à esquerda e à palavra mais à direita de cada expressão relevante multipalavra RE_i, ou seja, $lmostw(RE_i)$ e $rmostw(RE_i)$. $p(W,d_i)$ é a probabilidade da palavra W no documento d_i calculável através da equação (2.4).

$$p(W,.) = \frac{1}{\|D\|} \sum_{d_i \in D} p(W,d_i) \ , \quad (2.9)$$

Ao proporem a utilização de $Rvar(W)$, ou autores [1] tiveram como objectivo medir a variação da probabilidade da palavra W ao longo de todos os documentos da colecção.

Segundo os autores [1], a forma mais comum *Rvar*, de *Relative Variance*, é uma medida de variância ponderada, que é o segundo momento relativamente à média, e

[11] Ao realizar o presente trabalho não pude constatar esta afirmação.

que beneficia erradamente palavras muito frequentes sem significado semântico, como "de", "das", "e","ou", entre outras. Como os autores mencionam, isto acontece porque a diferença absoluta entre as probabilidades de ocorrência destas palavras ao longo de todos os documentos é alta, independentemente do facto de que geralmente ocorrem sempre em todos os documentos. Assim, estas diferenças são capturadas e sobrevalorizadas pela variância que mede o valor médio da quantidade (*distância à média*)^2 ignorando a ordem de magnitude das probabilidades individuais.

Para ultrapassar este problema, os autores introduziram uma alteração na fórmula de calcular a variância dividindo cada distância individual pela ordem de magnitude dessas probabilidades, ou seja, a probabilidade média, dado por $p(W,.)$ ver equações 2.7 e 2.8.

Resumindo, $Rvar(.)$ (Variância Relativa) na equação 2.5 reflecte essa alteração se for comparada à formula normal da variância que pode ser vista na fórmula da $Rvar(.)$ se se apagar $p(W,.)$ do denominador.

Assim, $LeastRVar(RE_i)$ é dado pelo menor valor $Rvar(W)$ considerando a palavra mais à esquerda e a palavra mais à direita de RE_i. Desta forma, os autores tentaram privilegiar as expressões relevantes mais informativas e penalizar as expressões multipalavras que contenham palavras sem significado semântico que iniciem ou terminem multipalavras extraídas automaticamente do tipo "relativamente a", "no que se refere a", etc.

Os autores de [1], partindo da observação de que geralmente a maioria das palavras sem significado semântico são geralmente curtas, de poucos caracteres, e de que, de um modo geral, palavras de maior comprimento têm uma maior acutilância semântica, introduziram também uma medida alternativa $LeastRVarLen$, definida em (2.10), que leva em consideração este aspecto.

$$LeastRVarLen(RE_i) = leastRVar(RE_i) . avgLen(RE_i) \quad , \qquad (2.10)$$

onde $avgLen(RE_i)$ é o comprimento médio de cada palavra da expressão RE_i, ou seja, número médio de caracteres de cada palavra de RE_i.

No trabalho realizado, de forma a se conseguir ter uma escalabilidade comparável nos resultados das avaliações, com a medida $Rvar(W)$, dada pela equação (2.8), optou-se

por harmonizar este valor dividindo pelo número total de documentos – 1, utilizando a equação,

$$Rvar(W) = \frac{1}{\|D - 1\|} \sum_{d_i \in D} \left(\frac{p(W, d_i) - p(W, .)}{p(W, .)} \right)^2 \quad , \quad (2.11)$$

Dos mesmos autores, temos em [7] uma variante desta medida, que em vez de utilizarem a média do comprimento das palavras, optaram por utilizar a mediana. Assim, definiram "*Pseudo Number of Words*", como

$$Pnw(MWE_i) = \frac{NumChars(MWE_i)}{Med(MWE_i)} \quad (2.12)$$

Onde, $NumChars(MWE_i)$ é o número de caracteres presentes na unidade multipalavra. E $Med(MWE_i)$ é a mediana do comprimento das palavras que compõem a unidade multipalavra em questão.

$$Cklen\ (MWE_i) = \frac{1}{|Pnw(MWE_i) - T| + 1} \quad (2.13)$$

Onde T é o número "típico" de palavras que uma palavra-chave tem. O valor máximo que $Cklen\ (MWE_i)$ atinge é um, se $Pnw(MWE_i)$ for igual a T. Tendo disponível estes valores, os autores em [7] improvisaram o $LeastRvar(MWE_i)$, obtendo a seguinte equação:

$$Mk(MWE_i) = LeastRvar(MWE_i) * Med(MWE_i) * Cklen(MWE_i) \quad (2.14)$$

Onde, segundo os autores, $Mk(.)$, privilegia unidades multipalavra que tenham não só as palavras mais à direita e mais à esquerda mais informativas, mas tendo também em conta palavras longas e um "*Pseudo Number of Words*" próximo de número "típico" de palavras que uma palavra-chave tem.

No trabalho realizado na elaboração desta dissertação, também foram criadas e utilizadas variantes de medidas que recorrem ao uso da Mediana (ver secção 3.2.4 da capítulo 3). A utilização da mediana, pelos resultados obtidos (ver capítulo 4), apesar

de mostrar por vezes alguns resultados interessantes, não tem o mesmo impacto que se verificou com a utilização do *Tf-Idf* e *Phi-Square*.

2.3.1.3 Chi Square

Esta métrica é muito utilizada na área de selecção de características para classificação, baseia-se num método probabilístico que interpreta um evento num conjunto de documentos, e dessa forma calcula o grau de ligação de uma característica a uma classe ou, no caso que investigarei, a um documento. Na equação seguinte $\chi^2(t,d)$ mede o valor da ligação do termo t ao documento d.

$$\chi^2(t,d) = \frac{N.(AD-CB)^2}{(A+C).(B+D).(A+B).(C+D)}, \qquad (2.15)$$

A - o número de vezes que o termo t e o documento d co-ocorrem;
B - o número de vezes que o termo t ocorre sem ser no documento d;
C - o número de vezes que o documento d ocorre sem o termo t;
D - o número de vezes que nem o documento d, nem o termo t ocorrem;
e **N** - o número total de documentos.

No cálculo da importância dos termos utilizando esta medida os termos que são mais negativamente relevantes para um documento não são ignorados. A medida calcula a frequência da presença e da ausência de um termo num documento e na colecção.

No trabalho realizado, comprovou-se por resultados obtidos que juntamente com o a medida *Tf-Idf*, o Chi-Square é das medidas utilizadas que melhores resultados produz na extracção de termos chave. Com o elaborar e desenrolar do trabalho, também vimos que quando analisamos os resultados para palavras e multipalavras juntas, o Chi-Square continuou a produzir resultados bastante bons. Ver capítulos 3 e 4 para mais detalhes.

2.3.1.4 Phi Square

O Phi - Square é uma variante do Chi – Square, e é dada pela expressão

$$\varphi^2 = \frac{\chi^2}{N}, \qquad (2.16)$$

Onde, N é o número total de termos presente no corpus, ou seja o somatório dos termos de todos documentos, e χ^2 o valor obtido na aplicação da equação (2.11). Esta medida foi utilizada com o objectivo de normalizar os resultados obtidos pelo χ^2.

No trabalho realizado, apesar de termos também trabalhado com o *Chi-Square*, os resultados obtidos em termos de ordenação das palavras, multipalavras e prefixos por grau da sua importância eram iguais aos resultados do *Phi-Square*. Optámos assim por fazer a avaliação final apenas com base no *Phi-Square* e por apresentar só esses resultados.

2.3.1.5 Informação Mútua

A métrica Informação Mútua [22], é bastante utilizada na modelação de linguagem e visa identificar associações entre termos aleatoriamente escolhidos, e nesse processo determinar a dependência que esses termos têm entre si. É calculada da seguinte forma,

$$I(t,c) = \log \frac{P_r(t \cap c)}{P_r(t)P_r(c)} \ , \qquad (2.17)$$

Onde, t é um termo e c a classe, no trabalho que realizei c representa o documento onde t ocorre. Segundo o trabalho de Filipa Madureira [5], "*esta expressão pode ser traduzida para o contexto da categorização de textos da seguinte forma".*

$$l(t,c) \approx \log \frac{A.N}{(A+C)(A+B)} \ , \qquad (2.18)$$

Onde, A representa o número de vezes que o termo t e a classe c co-ocorrem; B representa o número de vezes que o termo t ocorre sem ser na classe c; C representa o número de vezes que a classe c ocorre sem o termo t; e N representa o número total de documentos.

2.3.2 Não Estatísticas

Nesta secção descrevem-se outro tipo de metodologias de extracção que não recorrem a medidas estatísticas. Um exemplo é o trabalho apresentado em [23] onde a autora compara a utilização de gramáticas, de dois tipos, *"stochastic context-free grammar (SCFG)"*

e *"non-statistical context free grammar (CFG)"*, utilizando etiquetas morfo-sintáticas, de modo a conseguir extrair sequências de nomes e adjectivos (unigramas e bigramas).

Sendo uma CFG definida por uma gramática formal definida por uma quádruplo **G =** <**V, T, S, P**>. Onde **V** representa o conjunto de símbolos não terminais, **T** representa o alfabeto (o conjunto de símbolos terminais), **S** representa a categoria frase e **P** representa um conjunto finito de regras. A forma genérica dessas regras é apresentada na Figura 2.5.

$$X \longrightarrow \omega$$

Figura 2.5 – Regra de uma CFG

Exemplo retirado de [23]

Onde, X é um símbolo não terminal e ω é uma sequencia de terminais, T e não terminais V, como se exemplifica na Figura 2.6.

np	→	det	np
np	→	noun	noun
pps	→	prep	np
prep	→	*in*	
det	→	*the*	
noun	→	*DMA*	
verb	→	*controller*	

Figura 2.6 – "Simple Context Free Grammer"

Exemplo retirado de [23]

Onde *np* denota um sintagma nominal (*"noun phrase"*); *det,* um determinante, como é o caso do *"the; noun,* um nome como é o caso de *"DMA"; pps,* um sintagma proposicional (*"prepositional phrase"*); *prep,* uma preposição como é o caso de *"in"*.

Já uma *"stochastic context-free grammar (SCFG)"* é também uma gramática definida como um quádruplo como o anterior, mas com a diferenciação nas regras, que têm associado uma probabilidade, como se vê no seguinte exemplo,

$$(\quad X \longrightarrow \omega \quad , \quad p \quad)$$

Figura 2.7 - Regra de uma SCFG

Exemplo retirado de [23]

Onde, X é um símbolo não terminal e ω é uma sequência de terminais, T e não terminais V e p é a probabilidade da regra.

S	→	np	vp	.	0.193518
adjs	→	adv	adv	adj	0.0036083
np	→	noun	adj		3.20769e-05
np	→	noun	nounp		0.0256375
pps	→	prep	np	noun	1.05608e-05

Figura 2.8 – "Stochastic Context-Free Grammar"

Exemplo retirado de [23]

Onde *nounp* denota um nome próprio como seria o caso de *"DMA"* atrás.

No trabalho [24] o domínio das notícias é abordado, também numa perspectiva explicitamente declarada pelos autores de usar ferramentas linguísticas para extrair automaticamente palavras-chave. Estas ferramentas vão desde etiquetadores e analisadores morfológicos para as línguas que os autores querem usar, "*stemmers*"[12] para reduzir palavras que diferem apenas pelos seus sufixos a um radical comum. Usam ainda etiquetadores morfo-sintáticos para identificar padrões de etiquetas de palavras em queries e em documentos, como por exemplo a etiqueta NN (Noun[13], Noun). Utilizam também analisadores sintácticos ou segmentadores para identificar elementos frásicos ou multipalavras, e ainda léxicos semânticos[14] e heurísticas para reconhecimento de entidades com nome[15]. A utilização destas ferramentas, tornam obviamente este trabalho extremamente dependente da língua dos documentos a tratar. Apesar disso, os autores definem palavras-chave como sendo uma palavra simples, provavelmente nomes, ou multipalavras.

Existem outras metodologias, que estão geralmente associadas a ontologias, estejam estas especificadas à partida, ou não, sendo o seu principal objectivo obter um modelo representativo do domínio específico em questão. Podemos ver o trabalho realizado em [25] o qual permite proceder à análise de emails que tenham a proveniência de contactos não conhecidos e daí marcar esses emails como fraude ou não. Estas abordagens são bastantes limitativas, visto dependerem de uma ontologia que na maioria das vezes é específica a um domínio, impossibilitando o seu uso generalizado.

[12] Reduzir aos radicais
[13] Substantivo ou nome
[14] http://www.illc.uva.nl/EuroWordNet/
[15] Named entities

O grande problema associado às abordagens não estatísticas prende-se com o facto de que na maior parte dos casos exigirem a utilização de algo externo ao próprio texto que se esteja a analisar, nomeadamente gramáticas ou etiquetadores morfo-sintácticos. Desta forma, as abordagens não estatísticas são extremamente dependentes de uma língua ou de contextos muito específicos, não sendo fácil a sua adaptação para outras línguas ou a situações muito diferentes.

2.3.3 Híbridas

Por fim temos uma categoria, em que existe uma mistura que utiliza abordagens estatísticas e abordagens não estatísticas, como forma de se complementarem, ou seja, combina-se processamento estatístico com recurso a modelação linguística. Veja-se por exemplo [26], onde os autores utilizam gramáticas probabilísticas independentes do contexto em conjunção com métodos estatísticos. Lendo as palavras da autora "*adding linguistic knowledge to the representation (such as syntactic features), rather than relying only on statistics (such as term frequency and ngrams)*", identifica-se claramente o objectivo deste trabalho de não se basear somente nas estatísticas mas utilizar também conhecimento linguístico para melhorara extracção. Neste trabalho a autora realiza experiências com n-gramas, sintagmas nominais e com termos que coincidam com algum conjunto fixo de sequências de etiquetas morfo-sintáticas. Utilizou quatro características diferentes, frequência de termos, frequência dos documentos na colecção, posição relativa da primeira ocorrência e etiquetas morfo-sintáticas associadas com o termo. Este trabalho tem como objectivo o tratamento da extracção automática de termos chave como uma tarefa de aprendizagem automática, mais especificamente de classificação, o que implica que os autores treinem um classificador utilizando documentos com termos chave já conhecidos.

2.4 Extracção de Palavras

Na tese de Mestrado de Ventura [27] aborda-se a extracção de palavras (em oposição a multipalavras) relevantes, onde o autor cria duas métricas, a primeira denominada por *Score* que é uma medida estatística, para atribuição de relevância a palavras e baseia-se na análise da vizinhança das palavras. Esta medida baseia-se em duas componentes distintas, onde a primeira componente mede a importância de uma palavra num

determinado corpus baseado no estudo da relação entre essa palavra e as palavras que lhe sucedem imediatamente no texto. O *Score* do sucessor de uma palavra w, $S_{C_{suc}}(w)$, é calculada utilizando a equação seguinte

$$S_{C_{suc}}(w) = \sqrt{\frac{1}{\|Y\| - 1} \sum_{y_i \in Y} \left(\frac{p(w, y_i) - p(w, .)}{p(w, .)} \right)^2} \qquad (2.19)$$

Onde $\|Y\|$ representa o número de palavras distintas no corpus; e $p(w, y_i)$ representa a probabilidade de y_i ser um sucessor da palavra w; $p(w, .)$ representa a probabilidade média dos possíveis sucessores de w, que é dada por:

$$p(w, .) = \frac{1}{\|Y\|} \sum_{y_i \in Y} p(w, y_i) \qquad (2.20)$$

Onde,

$$p(w, y_i) = \frac{f(w, y_i)}{N} \qquad (2.21)$$

Onde N representa o número total de palavras no corpus e $f(w, y_i)$ é a frequência de ocorrência do bigrama(w, y_i) no mesmo corpus. Assim, esta componente, mede a variação da "*preferência*" da palavra w em ocorrer antes das restantes palavras do corpus.

Esta medida é uma variação da medida Rvar (secção 2.3.1.2) aplicada às palavras e às palavras que ocorrem imediatamente a seguir às palavras consideradas. É uma medida que pretendeu de certo modo, ultrapassar a impossibilidade de o LocalMaxs com o SCP (secção 2.2) extrair palavras relevantes no sentido de as multi-apalavras extraídas serem então designadas por expressões relevantes.

A segunda componente mede a *"preferência"* que uma palavra w tem para com as palavras que a antecedem, esta componente é designada por *Score* do antecessor ou $S_{C_{ant}}(w)$.

$$S_{C_{ant}}(w) = \sqrt{\frac{1}{\|Y\| - 1} \sum_{y_i \in Y} \left(\frac{p(y_i, w) - p(., w)}{p(., w)} \right)^2} \qquad (2.22)$$

Recorrendo às expressões (2.19) e (2.22), obtém-se o *Score* da palavra w, $S_c(w)$

$$S_c(w) = \frac{S_{C_{ant}}(w) + S_{C_{suc}}(w)}{2} \qquad (2.23)$$

Onde, através da média aritmética, se obtém uma métrica que permite classificar a relevância de uma palavra baseando-se nos resultados dos antecessores e sucessores dessa mesma palavra. Pelas expressões anteriores (2.19) e (2.22), e segundo o autor, a medida *Score* atribui maior valor a uma palavra quando esta tem tendência para se ligar a um conjunto restrito de palavras antecessoras e sucessores.

A segunda métrica que Ventura apresenta no seu trabalho é denominada por *Successor-Predecessor Quotient* (SPQ), que premeia as palavras que têm um maior número de sucessores e um menor número de antecessores, e é fornecida pela seguinte equação

$$SPQ(w) = \frac{\mathrm{N}\, suc(w)}{\mathrm{N}\, ant(w)} \qquad (2.24)$$

onde $\mathrm{N}\, suc(w)$ e $\mathrm{N}\, ant(w)$ representam respectivamente o número de sucessores distintos da palavra w e o número de antecessores distintos de w. Desta forma, segundo o autor, $SPQ(w)$ premeia as palavras que têm um maior número de sucessores e um menor número de antecessores, como é o caso dos nomes.

Neste mesmo trabalho, o autor, criou também o denominado Método das Ilhas que permite avaliar a relevância booleana de cada palavra com base em atributos estatísticos das palavras que ocorrem na vizinhança dessa mesma palavra. E que é considerado relevante se for tão ou mais relevante que todas as palavras que ocorrem na sua vizinhança imediata.

O trabalho desenvolvido que descrevo nesta dissertação, ao contrário de Ventura, não dá mais importância a uma palavra pela importância das palavras vizinhas, mas somente pela importância da própria no documento, eventualmente na colecção (no caso do *Phi-Square*, do *Rvar* e da Informação Mútua) mediante a aplicação de

medidas estatísticas, ver secção 2.3.1, ou das alternativas que foram desenvolvidas, que podem ser vistas na secção 3.2 do capítulo 3.

Num trabalho já referido anteriormente [26], foram feitas experiências também na extracção de unigramas relevantes, mas seguindo a metodologia descrita na secção 2.3.3.

David Ferreira, no seu trabalho [12], embora o seu objectivo fosse o de fazer *Clustering*[16] *de* Web Snippets, acabou também por medir a importância das palavras para descrever o conteúdo desses Web Snippets. A descrição do que foi feito pode ser vista em mais pormenor na secção 2.6.1.

Já o trabalho de Matsuo e de Ishizuka [28] também se enquadra na área de extracção de termos, mas a partir de um único documento. Estes autores que têm como objectivo apresentar um algoritmo de extracção de palavras-chave, neste caso, palavras ou sequências de palavras (bigramas), sem a utilização de um corpus. O algoritmo que os autores apresentam é descrito da seguinte forma: primeiro são extraídos os termos frequentes; de seguida as co-ocorrências de um termo com os termos mais frequentes são contabilizadas, preenchendo para isso uma matriz de co-ocorrências de termos par a par[17]. Este processo, repito, é feito para um único documento. Se um termo aparece frequentemente com um subconjunto particular termos, então esse termo aparenta ter importância. Assumindo que um termo w, aparece independentemente de termos frequentes, a distribuição de co-ocorrências do termo w e dos termos frequentes é similar à distribuição incondicional de ocorrências dos termos frequentes. Os autores, dividem um documento em frases, utilizando para isso possíveis separadores como "." ou "!" ou "?".

Para os autores, se um determinado termo w tem uma relação com um subconjunto particular de termos $g \in G$ dos termos frequentes, as co-ocorrências do termo w e g são maiores que o esperado, de onde se diz que a distribuição tem um desvio[18]. Assim, para os autores, um termo cuja co-ocorrência tenha um desvio, pode ter importância no documento. Por essa razão os autores usam o grau de desvio como um indicador de

[16] Agrupamento
[17] Paiwise term co-occurrences.
[18] Biased.

importância de um termo. O grau de desvio[19] da distribuição da co-ocorrência é calculada pelo uso da medida ao *Chi-Square (χ^2)*,

$$\chi^2\ (w) = \sum_{g \in G} \frac{\left(freq(w,g) - n_w p_g\right)^2}{n_w p_g}\ , \qquad (2.25)$$

Onde, w é o termo a testar, e $g \in G$ representa um conjunto de termos frequentes, e $n_w p_g$ representa a frequência esperada da co-ocorrência, e $\left(freq(w,g) - n_w p_g\right)$, representa a diferença entre as frequências esperadas e observadas. n_w é o numero total de termos nas frases em que w aparece. p_g é a soma do total de termos nas frases em que g aparece a dividir pelo número de termos no documento. Para os autores, um grande valor de $\chi^2\ (w)$ indica que a co-ocorrência do termo w mostra uma desvio grande. Os autores usam esta medida como um índice de desvios e não para testar hipóteses. Este trabalho, com um menor grau de satisfação lembra o trabalho de Ventura [27].

Uma outra forma de abordar a extracção de palavras recorrer a uma rede neuronal artificial [29], que é um modelo de programação que pretende ter semelhanças ao modelo neuronal biológico. Consiste num grupo de neurónios artificiais que processam a informação e a passam para outros neurónios artificiais. A ligação entre os neurónios permite formar uma rede complexa de grande poder computacional. O trabalho [30] é um exemplo da utilização de redes neuronais para a extracção de unigramas relevantes. Neste caso, cada nó da rede tem uma palavra associada aos termos pesquisados por um utilizador, com o mesmo peso inicial. Posteriormente recebe como entrada no modelo da rede, um documento, e se houver uma relação entre o documento e uma palavra presente nalgum dos nós, o peso desse nó é elevado a um nível superior. Esse peso tem como base uma "energia" que resulta da posição da palavra no documento. Este processo de evolução da rede neuronal continua até que seja alcançado um nível de estabilização de energia entre os nós, e o grupo de nós que tenha mais "energia" dá o valor de relevância desse documento associado às palavras procuradas.

[19] bias

2.5 Extracção de Multipalavras

Já nos referimos noutras secções deste capítulo, a trabalhos que fazem a extracção de multipalavras, nomeadamente [1] [10, 31]. Em qualquer destes trabalhos a extracção de multipalavras visa tão só este objectivo. Não pretendem extrair multipalavras que sejam necessariamente descritores do conteúdo dos documentos onde ocorrem.

Em [1] Joaquim F. da Silva et.al., utilizam o SCP, que é aplicado a um bigrama e é definido como se segue:

$$SCP(x,y) = p(x|y) * p(y|x) = \frac{p(x\,y)}{p(y)} * \frac{p(x\,y)}{p(x)} = \frac{p(x\,y)^2}{p(x) * p(y)}, \qquad (2.1)$$

Onde $p(x\,y)$, $p(x)$ e $p(y)$ são as probabilidades de ocorrência do bigrama (x y) e dos unigramas x e y no corpus; $p(x|y)$ representa a probabilidade condicional de x ocorrer à esquerda no bigrama (x,y) dado que y aparece à direita do mesmo bigrama.
Da mesma forma $p(y|x)$ representa a probabilidade de ocorrência de y ocorrer à direita no bigrama (x,y) dado que x aparece à esquerda no mesmo bigrama.
No entanto, a fim de se medir o valor de coesão de cada n-grama de um qualquer tamanho que possa aparecer no corpus, a normalização FDPN (*Fair Dispersion Point Normalizarion*) foi aplicada ao resultado da aplicação do SCP (.), por forma aos autores terem acesso a uma nova medida de coesão, denominada SCP_f (f de *"fair"*), esta medida está definida na equação (2.2).

$$SCPf\ (w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{\frac{1}{n-1} * \sum_{i=1}^{n-1} p(w_1 \dots w_i) * p(w_{i+1} \dots w_n)}, \qquad (2.2)$$

Onde, $p(w_1 \dots w_n)$, é a probabilidade do n-grama $w_1 \dots w_n$ ocorrer no corpus.

A ideia subjacente a esta fórmula é a de que é possível transformar qualquer n-grama de comprimento variável num pseudo bigrama, sendo que o pseudo-bigrama reflete parcialmente a coesão média entre quaisquer dois sub-n-gramas adjacentes contíguos em que foi partido o n-grama original. Daí o denominador de (2.2) ser a média de todos os produtos das probabilidades das partes em que foi dividido o n-grama.

O algoritmo LocalMaxs pode ser utilizado para extrair padrões de outros elementos dos textos além de expressões relevantes compostas por palavras, designadamente por caracteres ou por etiquetas morfo-sintácticas. Assim o algoritmo baseia-se na ideia de

que cada n-grama[20], e diz que entre cada n-grama existe uma espécie de "cola" ou coesão, que faz com que as palavras do n-grama fiquem juntas, e é definido como se segue:

Seja $W = w_1 \dots w_n$ um n-grama e $g(.)$ uma função de coesão genérica. E seja $\Omega_{n-1}(W)$ o conjunto de valores de coesão $g(.)$ para todos os $(n-1)$-gramas contíguos contidos no n-grama W. Seja $\Omega_{n+1}(W)$ o conjunto de valores de coesão $g(.)$ para todos os $(n+1)$-gramas contíguos que contenham o n-grama W. Seja, $len(W)$ o comprimento (número de elementos) do n-grama W.

Então, W é uma unidade multi Elemento (MEU) se e só se:

$$\forall_x \in \Omega_{n-1}(W),\ \forall_y \in \Omega_{n+1}(W)$$

$$(len(W) = 2 \wedge g(W) > y) \cup \left(len(W) > 2 \wedge g(W) > \frac{x+y}{2}\right)$$

Então, para n-gramas com n $\geq$ 3, o algoritmo elege todo o n-grama cujo valor de coesão seja maior que a média de dois máximos, o maior valor de coesão encontrado nos $(n-1)$-gramas contíguos contidos no n-grama W e o maior valor de coesão encontrado nos $(n+1)$-gramas contíguos que contenham o n-grama W.

Assim, no trabalho [1], o algoritmo LocalMaxs é utilizado como um extractor de multipalavras, onde os elementos MEU do LocalMaxs são vistos como sendo palavras.

Outro trabalho relacionado com a extracção de multipalavras, é o elaborado no artigo [32]. Aqui os autores apresentam um processo semi-automático para fazer sobressair recursos terminológicos num dado domínio específico. Os autores com o seu método visam processar linguisticamente texto "legível" pelos computadores e extrair uma lista de termos multipalavra candidatas, com a nuance de serem somente tratados bigramas, ou seja, multipalavras de duas palavras, que sejam representativas do domínio que se está a tratar, que posteriormente são validadas por peritos do domínio.

Os autores apresentam um método largamente baseado em análise linguística que se pode resumir aos seguintes passos. Primeiramente o texto é anotado morfo-sintáticamente tendo em conta o domínio do corpus. Este passo, contem duas componentes, um etiquetador morfo-sintático baseado num léxico morfológico e num

[20] Neste caso, um 1-grama é uma palavra, um 2-grama seriam 2 palavras e assim sucessivamente.

sistema que resolve ambiguidades morfológicas. O segundo passo é do de fazer o processamento do texto baseado numa gramática padrão para detecção expressões regulares e baseada em *"feature-structure Unification*", esta unificação, segundo os autores, é necessária para capturar concordância entre palavras *(e.g.* nomeadamente concordância de caso*)* na língua Grega. Por fim o resultado sofre uma lematização[21].
Como já referido, este método é baseado largamente no processamento e análise linguística do texto, onde posteriormente é aplicado uma análise estatística que serve para remover items resultantes do processo anterior que não apresentem evidência estatística suficiente para serem consideradas. Os trabalhos [10, 23, 24, 31] são exemplos deste tipo de abordagem à extracção de multipalavras.

Um outro trabalho, apresentado Ngomo em [33], apresenta uma metodologia só aplicável na extracção de multipalavras. Para tal, propõem uma nova métrica estatística denominada de SRE (*Smoothed relative expectation*),

$$SRE(w) = p(w)\frac{e^{-\frac{(d(w)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\frac{nf(w)}{\sum_{i=1}^{n} f(c_1 \dots c_i * c_{i+2} \dots c_n)} \qquad (2.26)$$

Onde, $d(w)$é o número de documentos onde w ocorre, μ e σ^2significam respectivamente a média e a variância da ocorrência de um n-grama num documento. $p(w)$ é a probabilidade de ocorrência de w no corpus. $f(w)$ é a frequência da ocorrência de w no corpus e $c_1 \dots c_i * c_{i+2} \dots c_n$ são padrões tais que a distância de Hamming de ham $(c_1 \dots c_i * c_{i+2} \dots c_n) = 1$.
Onde

$$ham(w, w') = \sum_{i=1}^{n} dif(w_i, w_i') \qquad (2.27)$$

com,

$$dif(w_i, w_i') = \begin{cases} 1 \; se \; w_i \neq w_i' \\ 0 \; caso \; contrário \end{cases}$$

[21] É o processo de agrupar as diferentes formas flexionadas duma palavra para que possam ser representadas por um único elemento (a forma singular no caso dos nomes, a forma masculina singular no caso dos adjectivos, e a forma infinitiva no caso dos verbos).

O autor fez a experimentação sobre o corpus TREC-9 para filtros adaptativos. Trata-se de um corpus composto por resumos (*"abstracts"*) de publicações do domínio da medicina. O autor fez comparações com outras medidas de extracção de multipalavras. O output do SRE foi uma lista ordenada de n-gramas dos quais η entre 100 e 10000 foram considerados em cada passo da avaliação. Na Figuras 2.9 e 2.10 podemos ver os resultados de Precisão e Cobertura documentados pelo autor no seu trabalho.

η	SRE
500	**29.40**
1000	**26.60**
1500	**26.26**
2000	**24.40**
2500	**22.96**
3000	**21.70**
3500	**21.45**
4000	**20.88**
4500	**20.31**
5000	**19.50**

Figura 2.9 – Precisão para a extracção de Unidades multipalavra**.**

Exemplo retirado de [33]

η	SRE
500	**1.05**
1000	**1.89**
1500	**2.80**
2000	**3.47**
2500	**4.08**
3000	**4.63**
3500	**5.34**
4000	**5.94**
4500	**6.50**
5000	**6.94**

Figura 2.10 - Cobertura para a extracção de Unidades multipalavra**.**

Exemplo retirado de [33]

A precisão da extracção de multipalavras descrita em [2] era de 81% para Português, 77% pata Inglês, 76% para Francês, 75% para Alemão e 73% para Português Medieval utilizando o SCP. Estes valores distinguem-se dos valores de precisão apontados na Figura 2.9.

O trabalho desenvolvido nesta tese, difere-se destes trabalhos vistos nesta secção já que tratamos multipalavras até ao nível de pentagramas (incluindo bigramas, trigramas, quadrigramas e pentagramas de palavras), configurável até mais se necessário. Além disso, como já dito anteriormente tratamos também palavras e prefixos de palavras.

2.6 Áreas de Possível Aplicação

Nesta secção apresentam-se alguns trabalhos em áreas onde a identificação da importância de termos relevantes faz parte de um processo mais complexo.

2.6.1 Agrupamento e Classificação de Documentos

Começamos pela área de Classificação e Agrupamento de documentos.
A Classificação de documentos é uma tarefa que consiste em atribuir um documento a uma ou a mais categorias, tendo como base para esta decisão o conteúdo desse mesmo documento tendo em linha de conta um conjunto de documentos. As tarefas de classificação de documentos podem ser divididas em dois tipos:

a) Classificação supervisionada, ou classificação propriamente dita, onde existe algum mecanismo externo, geralmente a interacção humana, para fornecer a informação sobre a classe (ou classes) a que o documento pertence. Na classificação propriamente dita, a colecção de documentos previamente classificados é dividido normalmente em dois conjuntos, um que vai servir para treinar um classificador e outro que vai servir para testar o grau de acerto do classificador previamente treinado na colecção de treino.
b) Classificação não supervisionada, ou agrupamento propriamente dito, onde a classificação/agrupamento deve ser feito sem suporte a nenhum mecanismo externo,

No que concerne à definição de agrupamento de documentos, podemos dizer que está intimamente relacionada ao conceito de agrupamento de dados. Agrupamento de documentos é uma técnica específica para a organização não supervisionada de documentos que envolve, extracção automática de tópicos, filtragens ou indexação rápida de informação. Mais, podemos afirmar que agrupamento de documentos e classificação de documentos envolve o uso de descritores, e de técnicas de extracção de descritores. Mas na classificação interessa-nos a ordenação das palavras e das

multipalavras e prefixos em termos da sua importância relativamente à classe e às classes, não relativamente aos documentos.

Sendo o principal objectivo do trabalho apresentado a ordenação de palavras-chave, através de medidas para a extracção de palavras e/ou multipalavras que sejam considerados como bons descritores de documentos, antevêmos uma possível futura utilização deste trabalho nas áreas de agrupamento e classificação de documentos, mais informação sobre esta discussão pode ser encontrada no capítulo 5.
Seguidamente apresentam-se alguns trabalhos realizados na área de agrupamento, em que se utilizam mecanismos para fazer a extracção de termos relevantes.

No trabalho desenvolvido por Fillippo Geraci *et al* [11], é-nos apresentado um problema de *Clustering*[22] de um conjunto de documentos num espaço de K grupos não sobrepostos, e apresentam um algoritmo escalável para *Clustering* de alta qualidade de Web Snippets[23]. A descrição do algoritmo sai fora do âmbito desta tese. Cada snippet é representado por um vector dos radicais das palavras do snippet. Para isso o snippet é pré-processado removendo-lhe palavras sem significado, reduzindo cada uma das outras palavras contidas no snippet aos seus radicais e por fim atribuindo pesos ("*cosine-normalized*" *Tf-Idf*) aos termos (radicais) obtidos. Para se ter uma ideia de como estes pesos são atribuídos ver secção 2.6.2, equação (2.33), em relação à sumarização de documentos.

Pensamos que o trabalho desenvolvido nesta tese, pode vir a ser aplicado em trabalhos futuros a realizar na área do agrupamento porque, como se verá mais tarde, faço uma análise comparativa entre várias métricas que poderão ser utilizadas para a atribuição de valores de peso a termos obtidos, neste caso particular, em Web Snippets ou a documentos.

Já no que diz respeito ao trabalho de Ferragina e Gulli [13] os autores apresentam um motor de pesquisa, SnakeT[24], que faz Agrupamento Hierárquico de Web Snippets. Ou seja, os autores pegam no resultado retornado por algum meta motor de pesquisa, e

[22] Agrupamento.
[23] Texto resultante de uma querie num motor de pesquisa, geralmente constituído por poucas dezenas de palavras.
[24] http://snaket.di.unipi.it/

apresentam o resultado dessa pesquisa numa hierarquia de *"directorias"* que são etiquetadas com elementos frásicos de comprimento variável, ver Figura 2.11.

Figura 2.11 -Resultado da query "Asthma"

Exemplo retirado de [13]

Este motor que nos é apresentado usa uma abordagem *"itemset-like"* para extrair etiquetas com significado, que capturam o tema dos snippets, contido na directoria em questão. A selecção/extracção das etiquetas é feita *on-the-fly*,a partir dos snippets vistos como *"gapped sentences"*, nomeadamente sequências de termos que ocorrem de forma não contígua, ou interrompida de comprimento variável. Sendo que a sua qualidade é enriquecida e avaliada recorrendo a duas bases de dados, uma resultante da indexação de uma colecção de textos âncora extraídos de mais de 200 milhões de páginas Web. Os textos âncora de uma hiperligação que aponte para uma página são utilizados, em tempo de execução, para enriquecer o conteúdo de snippets mais pobres de informação. Já a outra base de dados é um motor de ranking sobre uma directoria online, Dmoz.com, directoria esta que classifica mais de 3,500,000 sites em mais de 460,000 categorias. O motor de hierarquização utiliza o *Tf-Idf* sobre pares de palavras

que estão centradas nas categorias (*Dmoz - Categories*) presentes na base de dados do motor de ranking, ver Figura 2.12.

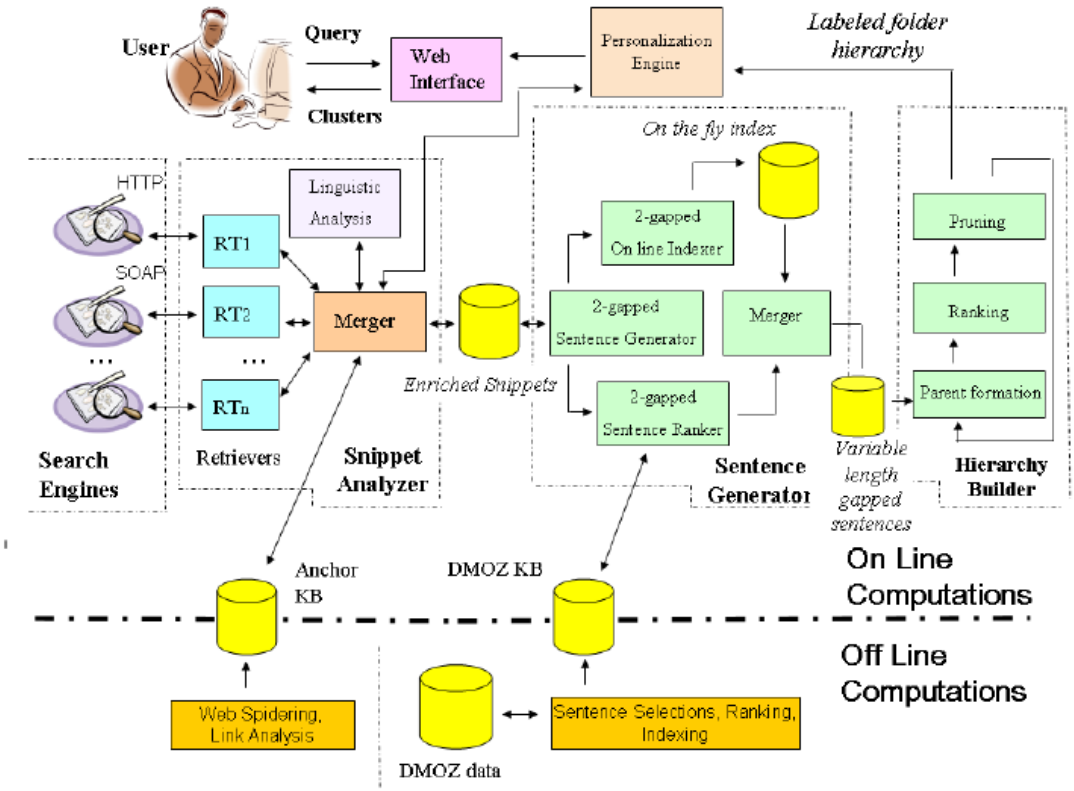


Figura 2.12 – Arquitectura do sistema Snaket,

Exemplo retirado de [13]

Ainda na área de clustering de web snippets, temos a tese de David Ferreira [12], onde o autor apresenta um trabalho onde o seu objectivo é o de fazer *Clustering* de Web-Snippets e a de propor a criação de "*uma nova abordagem"* para a criação de perfis dos utilizadores. Pretende fazer isso ao "*Criar os perfis dos utilizadores a partir da análise do histórico das pesquisas efectuadas pelos mesmos num motor de pesquisa capaz de efectuar categorização dos resultados.*". Para isso, criou modelos específicos por utilizador, e com isso construiu um "*sistema completamente autónomo e independente*" que parte de uma "*nova metodologia para efectuar a categorização de Web snippets baseada no cálculo do valor de importância das palavras*".

Para realizar este objectivo o autor decidiu "*Fazer uso das categorias associadas a cada pesquisa para extrair conhecimento oculto e auxiliar à criação dos perfis. Ou seja, em vez de analisar todos os documentos para extrair as categorias que mais*

sobressaem para um utilizador em questão, o sistema analisa a estrutura das queries bem como o conjunto de categorias que estão associados.". Assim, importa realçar a forma como o autor calcula o valor da importância das palavras. Primeiro extrai todas as características associadas a cada palavra ver a Tabela 2.1. O método de extracção das características sai fora do âmbito deste trabalho e não é descrito.

Característica	**Descrição**
W	Representação em caracteres da palavra
query_Word	Indicação se a palavra existe na query
F_Name	Número de ocorrências em que a palavra é considerada um nome
F_Acron	Número de ocorrências em que a palavra é considerada um acrónimo
S	Número de ocorrências em que a palavra está isolada
F	Frequência da Palavra
U	Número de Urls em que a palavra ocorre
WIL	Número de palavras contadas imediatamente à esquerda da palavra
WIR	Número de palavras contadas imediatamente à direita da palavra
WDL	Número de palavras diferentes contadas imediatamente à esquerda da palavra
WDR	Número de palavras diferentes contadas imediatamente à direita da palavra
W_Class	Raíz da palavra quando é utilizado *Stemming*.

Tabela 2.1 - Características analisadas numa palavra, tabela retirada de [12]

Tendo em conta estas características, em seguida calcula a importância de cada palavra recorrendo às seguintes propriedades:

***"Propriedade W1**: Se um termo aparece sozinho num segmento de texto, quer seja separado dos restantes termos por uma vírgula, um ponto ou outro separador, então é muito provável que esse termo tenha significado."*

$$W_1(w) = \frac{A(w)}{ln\big(F(w)\big)} \qquad (2.28)$$

Onde w é um qualquer termo, $A(w)$ é o número de vezes que w aparece sozinha e $F(w)$ é a frequência do termo w.

***"Propriedade W2**: Quanto maior for o número de termos que co-ocorrem com qualquer termo w tanto no contexto do lado esquerdo ou do lado direito, então menos importante esse termo será."*

$$W_2(w) = \frac{WIL(w) + WIR(w)}{2 * F(w)} \qquad (2.29)$$

Onde w é o termo, $WIL(w)$ e $WIR(w)$ são o número de termos que co-ocorrem nos lados esquerdo e direito do termo w e $F(w)$ é a frequência do termo w.

***"Propriedade W3**: Quanto maior for o número de termos diferentes que co-ocorrem com o termo w em ambos os seus lados esquerdo e direito comparativamente ao número total de termos existentes nos seus lados esquerdo e direito respectivamente, então, provavelmente menos importância terá essa palavra."*

$$W_3(w) = \left[\left(\frac{WDL(w)}{WIL(w)} + \frac{WDL(w)}{FH(w)}\right) * \frac{WIL(w)}{F(w)}\right] + \left[\left(\frac{WDR(w)}{WIR(w)} + \frac{WDR(w)}{FH(w)}\right) * \frac{WIR(w)}{F(w)}\right] \qquad (2.30)$$

Onde w é o termo, $WDL(w)$ e $WDR(w)$ são o número de termos diferentes que aparecem no lado esquerdo e direito do termo w. $FH(w)$ = Max $[F(w)]$, para todos os termos w. $WIL(w)$ e $WIR(w)$ são o número de termos que co-ocorrem nos lados esquerdo e direito do termo w e $F(w)$ é a frequência do termo w.

***"Propriedade W4:** Se um termo aparece designado pelo processo de pré-filtragem como sendo um nome ou um acrónimo com uma certa frequência num conjunto de texto, então é muito provável que esse termo tenha significado."*

$$W_4(w) = \frac{I(w)}{ln\big(F(w)\big)} \qquad (2.31)$$

Onde w é o termo, $I(w)$ é o valor da melhor representação do termos como sendo acrónimo ou nome e $F(w)$ é a frequência do termo w.

$$W(w) = \begin{cases} W_2(w) * W_3(w) \ , & se \quad W_1(w) < 0.5 \\ \dfrac{W_2(w) * W_3(w)}{1 + W_1(w)} \ , & se \quad W_1(w) \geq 0.5 \ e \ W_4(w) < 0.5 \\ \dfrac{W_2(w) * W_3(w)}{1 + W_1(w) + W_4(w)} \ , & se \ W_1(w) \geq 0.5 \ e \ W_4(w) \geq 0.5 \end{cases} \qquad (2.32)$$

Assim, baseado nestas quatro propriedades é possível ao autor atribuir um valor de importância $W(w)$ a um dado termo w e quanto mais baixo for esse valor, mais importante é o termo. Após ter esta etapa concluída o autor utiliza os resultados obtidos com as propriedades enumerada anteriormente para assim poder trabalhar com as palavras mais importantes encontradas, "*visto que estas representam um papel crucial no processo de categorização dos resultados"* para a nomear as categorias.

Para isso o autor utiliza um algoritmo que é executado em três passos. A criação dos pólos, onde é necessário ao autor inicializar o algoritmo para que sejam escolhidos os termos mais representativos. Com esse propósito, todas as palavras que se situem entre as primeiras posições da lista ordenada de palavras mais importantes para cada url e que existam em mais de dois urls, são propostas para centros iniciais de clusters (os ditos pólos), a unificação e absorção, escolha de um nome identificador para o conteúdo do cluster. A descrição do algoritmo sai foram do âmbito do trabalho realizado na presente tese, mas realça-se o último passo, em que através da união e absorção cada cluster pode conter mais do que uma potencial palavra para descrever o seu conteúdo, mas pode acontecer que os urls do cluster contenham outro tipo de palavras, nesta caso, multipalavras ou outro tipo de expressão composta que providenciem uma etiqueta mais interessante para o cluster. Da aplicação da equação (2.32) que fornece ao autor o grau de importância destas expressões, este valor é utilizado para fazer uma comparação entre as frequências das etiquetas simples que identificam o cluster. Caso o valor da expressão seja maior que um valor de proporcionalidade com a frequência da palavra etiqueta do cluster, então a expressão

composta é promovida a etiqueta do cluster, caso contrário a palavra simples mantém-se como etiqueta.

Uzun [34] aborda a extracção de palavras-chave que sejam palavras significantes de um documento, e considera esta problemática como sendo um problema de classificação. O método apresentado, para identificar as palavras-chave, utiliza um classificador "naive Bayesian", que utiliza o *Tf-Idf* para fornecer a pontuação da palavra, a distância da palavra em relação ao início do texto, do parágrafo e do frase. Assume que as características de uma palavra-chave têm uma distribuição normal e que as palavras-chave são independentes. O método segue uma linha de aprendizagem supervisionada, classificação, ao utilizar palavras-chave já extraídas de documentos presentes no corpo do conjunto de treino.

2.6.2 Sumarização de Documentos.

Com a quantidade de informação presente em documentos electrónicos, e com a tendência para o seu número aumentar cada vez mais, os métodos de sumarização de documentos são cada vez mais importantes.

No trabalho realizado por Marina Litvak e Mark Last [35], que exemplifica duas abordagens novas, uma supervisionada, logo uma abordagem de classificação, e outra não supervisionada vulgo agrupamento. Os autores neste trabalho apresentam o primeiro passo de extracção de sumários onde as palavras mais salientes ("palavras-chave") são extraídas para gerar o sumário. Como cada palavra distinta é representada como um nó do grafo do documento, os autores reduzem o problema de extracção de palavras-chave ao problema de extracção de nós salientes em grafos. Ou seja, as duas abordagens baseiam-se na representação sintáctica baseada em grafos que representam textos e documentos Web, onde os nós mais salientes dos grafos representam as palavras-chave dos documentos em causa. Esta representação em grafo, é definida como representando os arcos as relações entre palavras, e representando cada nó uma única palavra, ou seja, não há repetição de nós mas sim o incremento de um contador do número de vezes que essa palavra ocorre num nó que já exista.
Se uma palavra X precede imediatamente uma palavra Y na mesma frase algures num documento, então passa a existir um arco direccionado de X para Y.

Na abordagem supervisionada, de classificação, os autores para tentarem identificar os nós salientes do grafo treinaram algoritmos de classificação numa colecção de textos com o objectivo de induzir um modelo de identificação de palavras-chave. Cada nó de cada grafo de cada documento pertence a uma de duas classes, *"YES"* se a palavra correspondente está incluída no sumário extraído do documento, *"NO"* caso contrário. Os autores consideram características de um grafo, nomeadamente o grau[25] do nó, que caracteriza a estrutura do grafo bem como características estatísticas. AS características são as seguintes: *"In Degree"*[26], número de ligações que entram; *"Out Degree"*[27] número de ligações que saem; *"Degree"*, número total de ligações. A *"Frequência"* do termo representado pelo nó. A *"Distribuição das palavras frequentes"* valor ente zero e um, sendo 1 se a frequência do termo for maior ou igual a um limite[28]; o *"Location Score"* que calcula uma média de valores (*"Scores"*) de localização entre todos as frases que contenham a palavra N representada pelo nó; o "*Tf-Idf"* da palavra representada pelo nó; e o *"Headline Score"* valor ente zero e um, sendo um se e só se o título do documento contem a palavra representada pelo nó.

Na abordagem não supervisionada, de agrupamento, correram o algoritmo HITS no grafo do documento sob a assunção que os nós mais bem classificados devem representar as palavras-chave do documento. O algoritmo HITS é capaz de distinguir entre *"autoridades"* páginas com um grande número de links a entrar, e *"Hubs"* páginas com um grande número de links de saída. Para cada nó o HITS produz dois conjuntos de resultados. Um valor para "autoridade" e um valor para "hub".
A experimentação efectuada neste trabalho foi feita sobre uma colecção de sumários de referência. Dado um conjunto de documentos de treino, a classificação supervisionada fornece a identificação de palavras-chave mais certeira, enquanto a F-measure mais alta é alcançada com um simples *degree-based ranking*.
Na abordagem não supervisionada, é suficiente apenas executar a primeira iteração do HITS em vez de o executar em toda a sua convergência.

Em [36] os autores, abordam a questão da sumarização de documentos da Web tendo em conta o contexto dos mesmos. O contexto do documento Web é considerado como

[25] degree
[26] Número de Setas a entrar no nó.
[27] Número de setas a sair do nó.
[28] Para os autores este limite é de 0.05

sendo o conteúdo textual de todos os documentos que tenham uma ligação[29] ao documento em causa. Segundo os autores, a eficiência desta abordagem depende do tamanho do conteúdo e do contexto do documento alvo sobre o qual se trabalha. No entanto, sua eficiência depende também da existência de ligações[30] para os documentos de destino, sem deixar de ter em conta a quantidade e a qualidade dessas mesmas hiperligações.

Neste trabalho, os autores abordam as especificidades inerentes ao facto de se trabalhar na sumarização de documentos baseadas em contexto, nomeadamente a contextualização, a parcialidade[31] e a topicalidade[32]. Entende-se por contextualização a extracção de porções de informação entre os documentos do contexto que estão ligados ou têm informação sobre o documento alvo. Já por parcialidade podemos dizer que são os pedaços de informação partilhados pelos documentos do contexto que só dizem respeito a parte do conteúdo do documento alvo. Têm então de ser colocados juntos para que cubram inteiramente o alvo, ver no exemplo extraído de [36], "*cars robbed in Nevada*" seria uma parte importante do contexto.

> *"1. < LINK >CNN< /LINK > reported the rate of cars robbed in Nevada has increased of 5% in the second quarter.*

Entende-se por topicalidade a distinção que se tem de fazer entre os elementos que estão relacionados com o documento alvo, mas que não fornecem nenhuma pista sobre o conteúdo do documento alvo, como se pode ver no exemplo extraído de [36].

> *"2 < LINK >CNN< /LINK > is a news website. In the next sections, these issues will be discussed."*

Os autores começam por abordar o problema da contextualização, processo se refere a todos os passos intermédios necessários para juntar as frases do seu contexto. Decidiram usar um modelo baseado em vectores para representar estas frases. Este modelo usa vectores de termos pesados. Estes pesos resultaram do uso do *Tf-Idf*, dado pela seguinte equação:

[29] Link Web.
[30] Links Web.
[31] Partiality.
[32] topicality

$$w_{ik} = \frac{tf_{ik} * log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{j=1}^{N}\left(tf_{ij}\right)^2 * \left(log\left(\frac{N}{n_j}\right)\right)^2}} \qquad (2.33)$$

Onde, tf_{ik} é a frequência de ocorrência do termo *Wk* na frase *Si,* se tf_{ik} é zero se *Wk* não aparecer em *Si. N* é o tamanho do contexto e n_k é o numero de documentos no contexto com o termo *Wk.*

Seguidamente abordam a parcialidade, que é abordada pelos autores como sendo a extracção de *"representantes"* do contexto de um documento alvo. O conjunto de *"representantes"*de um contexto é o subconjunto mais pequeno de frases do contexto, que removendo um elemento, faria com a informação mais global do contexto decrescesse. Para saberem que frases podem remover até chegarem às frases "*representantes"* os autores definiram uma medida de inclusão que denominaram de *"inclusion measure", da*das duas frases S_i e S_k, o valor de inclusão $I(S_i, S_k)$, de S_i incluída em S_k, é definida como se segue:

$$I(S_i, S_k) = \frac{\sum_{j=1}^{N} w_j^i . w_j^k}{\sum_{j=1}^{N} w_j^i} \qquad (2.34)$$

Onde, as frases S_i e S_k, são representadas pelos vectores <w_1^i ... w_N^i > e <w_1^k ... w_N^k >.

Seja $S = \{S_i\}_{i=1\,..N}$ o contexto de um documento. As frases que podem ser removidas do contexto sem perca de informação são definidas pelo conjunto,

$$S' = \{S_i \ : \ \exists_k \ \neq \ i, \ I(S_i, S_k) = 1\}$$

Então, o conjunto de "*representantes"* é $S - S'$.

Finalmente abordam a topicalidade, que foi formalizado pelos autores da seguinte forma, uma "*frase de referência*"[33] é uma frase cujo conteúdo não contem qualquer pista sobre o conteúdo do alvo. E uma "*frase sujeito*"[34] corresponde a uma situação onde o conteúdo da frase dá uma boa ideia sobre o conteúdo do documento alvo. Isto

[33] Reference sentence
[34] Subject sentence

levou aos autores a definirem uma medida denominada como *"degree of topicality of a sentence S with a Document D"* que devolve um valor entre zero e um. Tal que:

T (S, D) = 0 significa que S é uma referência a D

T (S, D) = 1 significa que S é um assunto de D.

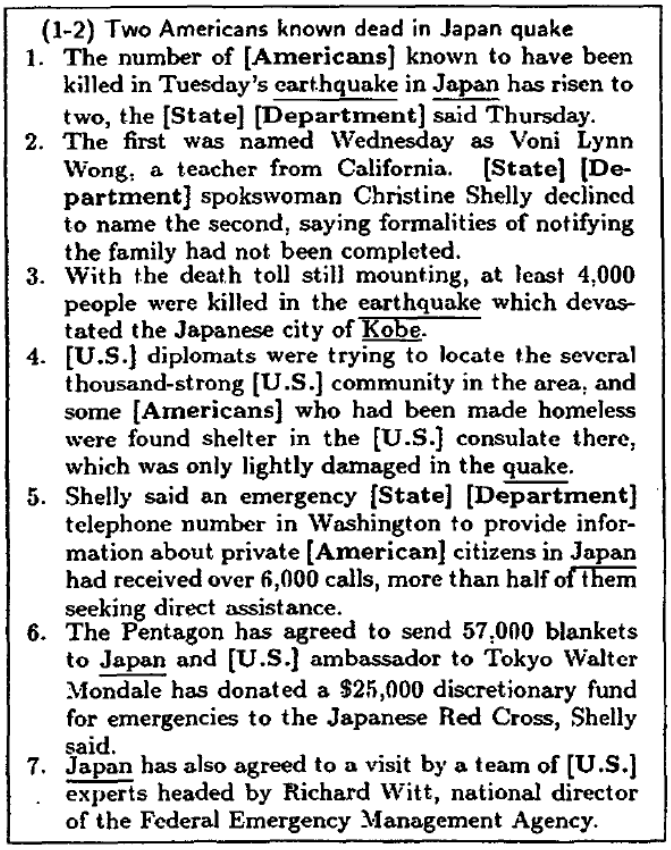
Onde T (S, D) dá como resultado um valor de satisfabilidade, é definido como se segue:

$$T(S, D) = \frac{|S \cap D|}{|D|}$$

Onde a intersecção de S com D, significa o grau de *topicality* de uma frase C com um documento D.

Neste ponto os autores indicam duas abordagens. Uma das abordagens leva em linha de conta tanto o conteúdo como o contexto do documento, enquanto a outra só tem em consideração os elementos do contexto do documento. Resumindo os autores recorrem à extracção das frases mais relevantes do documento a ser tratado, recorrendo ao uso da representação do documento como um vector de pesos de palavras calculada utilizando (2.33) recorrendo ao *Tf-Idf* normalizado e a uma medida de similaridade, de forma a produzir automaticamente um sumário, que pode conter não só o conteúdo principal, como pode incluir também outros conteúdos de vários tópicos diferentes.

Outro trabalho também incluído nesta categoria é apresentado em [14], que foi baseado em estudos preliminares reportados no relatório final [37] e que se baseia na identificação do tópico e do evento de cada documento, que são diferenciados pelos autores. Mas quer o tópico quer o evento são palavras. Os autores partem da assunção de que um evento associado a um documento aparece ao longo de vários parágrafos, enquanto um tópico não.



(1-2) Two Americans known dead in Japan quake

1. The number of [Americans] known to have been killed in Tuesday's earthquake in Japan has risen to two, the [State] [Department] said Thursday.
2. The first was named Wednesday as Voni Lynn Wong, a teacher from California. [State] [Department] spokswoman Christine Shelly declined to name the second, saying formalities of notifying the family had not been completed.
3. With the death toll still mounting, at least 4,000 people were killed in the earthquake which devastated the Japanese city of Kobe.
4. [U.S.] diplomats were trying to locate the several thousand-strong [U.S.] community in the area, and some [Americans] who had been made homeless were found shelter in the [U.S.] consulate there, which was only lightly damaged in the quake.
5. Shelly said an emergency [State] [Department] telephone number in Washington to provide information about private [American] citizens in Japan had received over 6,000 calls, more than half of them seeking direct assistance.
6. The Pentagon has agreed to send 57,000 blankets to Japan and [U.S.] ambassador to Tokyo Walter Mondale has donated a $25,000 discretionary fund for emergencies to the Japanese Red Cross, Shelly said.
7. Japan has also agreed to a visit by a team of [U.S.] experts headed by Richard Witt, national director of the Federal Emergency Management Agency.

Figura 2.13 - Um Documento intitulado "Two Americans dead in Japan quake",

Exemplo retirado de [14].

No texto da Figura 2.13, as palavras *"Japan"* e *"quake"* são palavras, tópico e evento em simultâneo. Os próprios autores admitem que esta diferenciação entre tópico e evento nem sempre se verifica, e podem também existir casos onde uma mesma palavra pode ser tópicos e evento ao mesmo tempo, segundo a definição dos autores.

Quando acontece uma colisão destas os autores, assumem a palavra como sendo um tópico e não um evento. Assim, os autores apresentam uma metodologia para extrair parágrafos chave com o objectivo da sumarização de multi-documentos, documentos de notícias difundidos por cadeias noticiosas, com base em tópicos e eventos. A técnica que os autores usam para fazer a distinção entre tópico e evento explora explicitamente a característica denominada por dependência do domínio das palavras[35], ou seja, o quão fortemente uma palavra caracteriza um conjunto de dados.
O método dos autores, assume que um evento associado a um documento aparece ao longo de parágrafos enquanto uma palavra tópico não. Assim, para efectuarem a extracção de tópicos e eventos, dividem esta tarefa em duas observações:

[35] "Domain Dependency of Words"

a) Se uma determinada palavra aparece ao longo de parágrafos (documentos);
b) Se uma palavra aparece ou não frequentemente.

A situação descrita em a) é representada por um valor de dispersão, dado pela equação (2.36) indicada abaixo, enquanto b) por um valor de desvio, dado pela equação (2.37), indicada abaixo.

A seguinte formulação é análoga no cenário em que se tratam documentos ou se tratam parágrafos. Assim, o primeiro passo do método dos autores é o de associar um peso a cada palavra individualmente num documento, e aplicaram a métrica *Tf-Idf* ao nível do documento (e ao nível de parágrafos).

$$Wd_{it} = TFd_{it} * log\frac{N}{Nd_t} \quad (2.35)$$

Onde, Wd_{it} é o valor de *Tf-Idf* de um termo t no n-ésimo documento i. A mesma fórmula é usada para calcular o peso das palavras nos documentos e nos parágrafos, bastando para isso substituir em (2.35) d (de documento) por p (de parágrafo). N é o numero de documentos e Nd_t o número de documentos onde o termo t ocorre.Para parágrafos N representa o número de parágrafos e substituindo Nd_t por Np_t temos o número de parágrafos onde t ocorre.

O segundo passo do método dos autores, é o de calcular a dependência do domínio das palavras, que é calculado recorrendo ás seguintes formulações:

$$DispD_t = \sqrt{\frac{\sum_{i=1}^{m}(Wd_{it} - mean_t)^2}{m}} \quad (2.36)$$

$$Devd_{it} = \frac{(Wd_{it} - mean_t)}{DispD_t} * 10 + 50 \quad (2.37)$$

Onde a equação (2.36) dá o valor da dispersão do termo t ao nível do documento da colecção de m documentos.

Da mesma forma, $DispP_t$ dá o valor da dispersão do termo t ao nível do parágrafo.Já a equação (2.37) denota o valor do desvio do termo t no n-ésimo documento.

Analogamente, $Devp_{it}$, denota o desvio do termo t no n-ésimo paragrafo. Em ambas as equações (2.36) e (2.37) $mean_t$ é a média do total dos valores de *Tf-Idf* do termo t ao nível de documento.

Tendo isto, o último passo do método dos autores é extrair as palavras que sejam tópico e as que sejam eventos, utilizando as equações (2.36) e (2.37).

Como muitos dos trabalhos apresentados, estes autores apenas se orientaram para o tratamento de palavras, ignorando multipalavras. Trabalham a dois níveis distintos, o de documento inteiro e ao nível de parágrafo.
No trabalho desenvolvido nesta dissertação, não tem esta a opção. Trabalha-se com o corpus total, e com os documentos em particular. Não pretendemos sumarizar no sentido de extrair frases ou parágrafos que de alguma forma representem o conteúdo de do documento. Pretende-se extrair palavras e multipalavras que representem o conteúdo do documento.

2.6.3 Construção de Ontologias

Uma ontologia [38] é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e as relações entre estes. É normalmente utilizada para fazer inferências sobre os objectos do domínio. Em particular, uma ontologia de domínio específico é uma ontologia que modela um determinado domínio, ou somente parte dele. Representa o significado particular de termos no respectivo domínio. Estes termos, mesmo sendo extraídos automaticamente, têm de ser sempre validados por um perito do domínio. Estes termos, geralmente são palavras, com relações entre si. Por exemplo, tomemos a palavra carta. Uma carta pode ter vários significados, uma ontologia sobre o domínio *poker* iria modelar a carta como uma carta de jogo, enquanto que uma ontologia sobre comunicação iria dar o significado de documento escrito de uma pessoa para outra.
O trabalho desenvolvido nesta tese, pode extrair os termos mais importantes de um conjunto de documentos de um determinado domínio, e fornecer a um perito do domínio uma forma mais prática de aceder a possíveis termos para enriquecer uma ontologia, ou os termos base para a criação de uma novo ontologia. Mas neste caso, teríamos de centrar a importância das palavras, multipalavras e prefixos relativamente aos termos que podem aparecer na meta-informação de um documento. Assumindo que esses termos de conteúdo são termos de uma ontologia de organização que modula esses documentos.

Já em [39] é apresentado um método para ajudar um *"Knowledge Enginner" a* identificar conceitos importantes num determinado domínio de uma ontologia. Que no trabalho do autor são palavras e multipalavras que transmitem um significado simples, ou complexo, dentro um determinado domínio a partir de documentos como páginas Web. O método baseia-se em duas medidas, Relevância do Domínio (*Domain Relevance*), DR, e Consenso do Domínio[36] (*Domain Consensus*), DC; que fornecem a especificidade de um termo candidato a termo do Domínio. Os autores sentiram esta necessidade, porque num texto existem termos que podem ser muito frequentes como "tempo real" ou "semana passada", mas que são pouco significativas em termos de descritibilidade dos conceitos do domínio. Por isso, os autores criaram a medida de Relevância do Domínio para testar a especificidade de um determinado candidato terminológico tendo em conta um determinado domínio. E definiram esta medida como

$$DR(t, D_i) = \frac{P\ (t|D_i)}{\sum_{i=1}^{n} P(t|D_i)} \qquad (2.38)$$

Onde D_i denota o domínio *i*, t um termo e sendo a Probabilidade condicional $P\ (t|D_i)$ estimada da seguinte forma:

$$E\big(P\ (t|D_i)\big) = \frac{freq\ (t\ em\ D_i)}{\sum_{i=1}^{n} freq\ (t\ em\ D_i)} \qquad (2.39)$$

Onde $E(\ )$, denota a estimativa da probabilidade.

Já o Consenso do Domínio é uma medida que mede a distribuição do uso de um termo num determinado domino D_i. Ou seja, a distribuição de um termo t ao longo de d_j documentos, pode ser visto como uma variável estocástica estimada através de todos os $d_j \in D_i$. A entropia H desta distribuição expressa o grau de consenso do termo t em D_i. O que, expresso numa fórmula, é visto como:

$$DC\ (t, D_i) = H\left(P(t, d_j)\right) = \sum_{d_j \in D_i} P(t, d_j)\ log_2\left(\frac{1}{P(t, d_j)}\right) \qquad (2.40)$$

Onde,

[36] "Domínios" são programaticamente representados por colecções de textos sobre diversas áreas, medicina, finanças, turismo, etc.

$$E\left(P(t,d_j)\right) = \frac{freq\left(t\ em\ d_j\right)}{\sum_{d_j \in D_i} freq\left(t\ em\ d_j\right)} \qquad (2.41)$$

Onde $E(\ )$, denota a estimação.

Já nos trabalhos [40-42] realizados por Fortuna *et. al.* onde entre outros avanços científicos se propõe a criação semi-automática de uma ontologia de tópicos. O Sistema apresentado pelos autores apresenta tópicos ao perito do domínio no momento em que este está a definir a ontologia. Para alcançar este objectivo, os autores no trabalho [42] usam duas técnicas para extrair tópicos de documentos: *Latent Semantic Indexing* e *K-Means Clustering.* Para começar os autores trabalham na representação de documentos, baseada num modelo vectorial onde os textos são transformados num saco de palavras ao mesmo tempo que são atribuídos pesos às palavras com recurso ao *Tf-Idf*. Referem ainda que a similaridade entre dois documentos é definida como o *coseno* do ângulo entre os seus vectores representantes (*cosine-smilarity*).

Tendo esta base os autores aplicam então a *Latent Semantic Indexing [41]* que é uma técnica para extrair *background Knowledge* a partir de documentos de texto. Usa uma técnica da álgebra linear denominada de SVD (*Singular Value Decomposition*) e um saco de palavras para detectar palavras com significados similares, o que segundo os autores também pode ser visto como a extracção de conceitos com semântica escondida ou tópicos de documentos. Em simultâneo também utilizam o *K-Means Clustering [41]*, para particionar dados com o objectivo de que cada C*luster* contenha apenas pontos que são similares de acordo com alguma métrica pré-definida. No contexto de texto isto pode ser visto como encontrar grupos de textos similares, ou seja, documentos que partilhem palavras similares.

Os autores usam dois métodos. O primeiro visa extrair tópicos utilizando vectores de centróides, sendo um centróide a média do somatório de todos os vectores dentro do tópico. E o segundo método baseia-se, segundo os autores, no trabalho de [43] utilizando o classificador binário *Support Vector Machines* [44]. A diferença na utilização destes dois métodos utilizados pelos autores é a de que uma leva em linha de conta o contexto do tópico enquanto que a outra não. Ambas diferem das medidas utilizadas nesta tese, apesar de "partilharem" um objectivo comum, o de encontrar palavras-chave.

2.6.4 Povoamento de Ontologias

Uma outra maneira de trabalhar com ontologias, é a de povoar as mesmas, ao invés de as construir de raiz. Nesta abordagem encontram-se trabalhos que se focam essencialmente em problemas de domínios específicos. Tome-se como exemplo o trabalho realizado em [45] onde os autores propõem uma metodologia para retirar informação pessoal de membros de um departamento da universidade, extrair informação composta pelo grau académico, email, número de telefone da pagina pessoal da pessoa em questão, identificação de grupos de pessoas que trabalhem juntas através da monitorização de listas de publicações, e em projectos de investigação que essas pessoas estejam envolvidas.

Como os autores não tinham classificadores disponíveis para usar, começaram por identificar os nomes de pessoas utilizando um NERC (Named Entety Recognizer), os autores não especificam no seu trabalho qual o NERC que utilizaram, sendo que os nomes identificados pelo NERC são ainda validados recorrendo a serviços como o CiteSeer (citeseer**.**com).

Outro trabalho é apresentado em [46] onde os autores descrevem o sistema artequakt[37], este sistema procura a Web e extrai informação ou conhecimento sobre artistas, baseado numa ontologia que descreve esse domínio, e posteriormente guarda esse conhecimento numa base de conhecimento que depois é usada para produzir biografias personalizadas de artistas.

[37] http://www.aktors.org/technologies/artequakt/

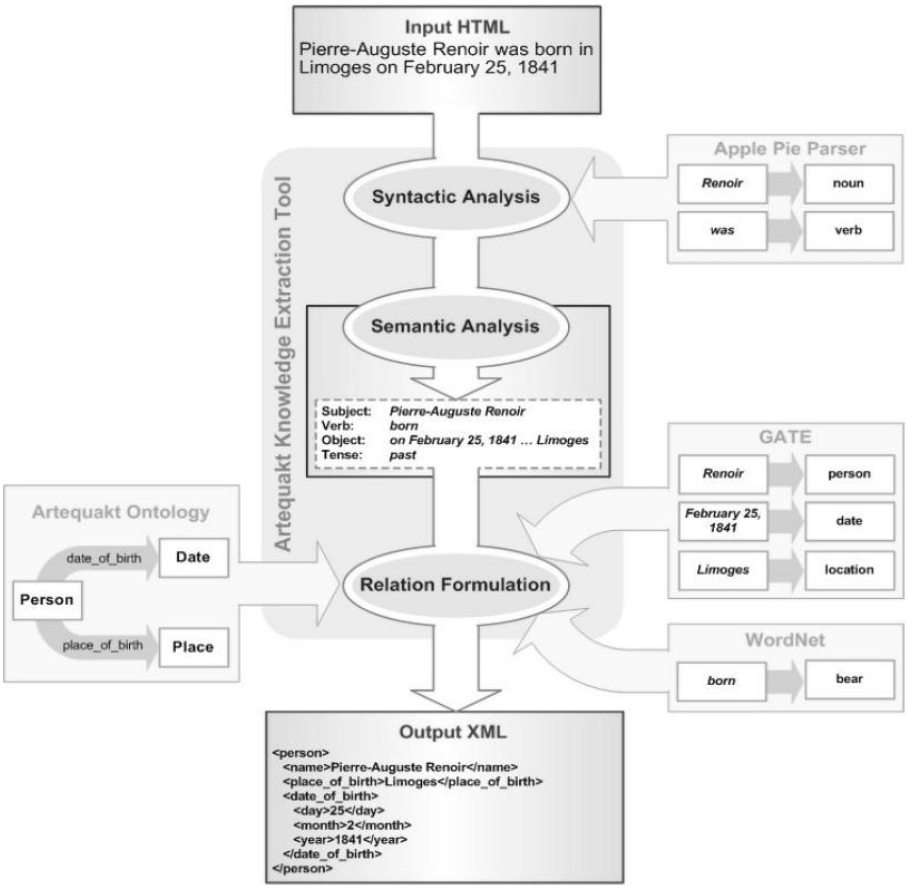


Figura 2.14- Processo de Extracção de Informação do Artequakt's,

Imagem retirada de [46]

A forma como o sistema de extracção de conhecimento dos autores funciona, é através da identificação e extracção de triplos de conhecimento[38] (conceito – relação – conceito) de documentos e fornece-os ao resto do sistema dos autores como ficheiros XML. Este processo é iniciado com a busca de documentos Web num qualquer motor de pesquisa, seguidamente este documento é processado para reconhecimento de entidades com nome. No caso deste trabalho os autores usam o sistema GATE[39]. Após este passo o procedimento de extracção é processado sendo cada documento dividido em parágrafos e em frases, onde cada frase é analisada sintáctica e semanticamente para extrair os triplos relevantes. Na análise sintáctica são extraídos grupos de palavras para funções sintácticas sem ter em consideração o seu significado semântico. Os autores fazem este processo recorrendo ao *"Apple Pie Parser"*[40]. Na análise semântica as frases são decompostas em frases mais simples para possibilitar a localização dos principais componentes como sujeitos, verbos e objectos, esta localização é

[38] Knowledge triplets
[39] http://gate.ac.uk/
[40] http://nlp.cs.nyu.edu/app/

conseguida pelo uso do GATE e do "highlight" dado pelo WordNet[41]. Que na frase seguinte faria sobressair *"Pierre-Auguste Renoir"* como o nome de uma pessoa, *"February 25, 1841"* como uma data e *"Limoges"* como um local.

O uso de informação lexical por parte dos autores, torna o trabalho mais dependente da língua dos documentos, visto os triplos poderem variar consoante a língua que se está tratar. Apresento a seguir um exemplo retirado do trabalho dos autores que ilustra o processo utilizado. Dada a seguinte frase:

"Pierre-Auguste Renoir was born in Limoges on February 25, 1841."

Seriam produzidas as seguintes relações ontológicas

<Pierre-Auguste Renoir> <date_of_birth> <25/2/1841>

<Pierre-Auguste Renoir> <place_of_birth> <Limoges>

2.7 Observações sobre as Áreas Possíveis de Aplicação

Como podemos ver na secção anterior, existe uma panóplia de aplicações onde a necessidade de se extrair palavras-chave é importante, independentemente de serem só palavras, ou multipalavras. O nosso objectivo nesta tese é o de trabalhar quer com palavras, quer com multipalavras acrescentando ainda o uso de prefixos de palavras. Esta opção deve-se ao facto de se pretender trabalhar também com línguas morfologicamente ricas. Por exemplo, em checo a palavra "mesa" se utilizada como sujeito tem uma forma, se for utilizada como complemento directo tem outra e se for considerada como o objecto indirecto ainda tem outra, para além da possibilidade de utilização de mais quatro casos, perfazendo sete no total.

E estas palavras posteriormente extraídas podem ser utilizadas como etiquetas de possíveis clusters, como descritores do conteúdo de documentos, como possíveis tópicos a serem incorporados numa ontologia.

2.8 Medidas de Avaliação de Resultados

2.8.1 Precision e Recall

A *Precision* e o *Recall* são duas medidas estatísticas, que trabalham com informação binária, e servem para avaliar a qualidade dos resultados obtidos em domínios tais como a Recuperação de Informação, Text Mining, Data Mining, etc.

[41] http://wordnet.princeton.edu/

As suas expressões são as seguintes:

$$Precision = \frac{\#\left(termos_{\ relevantes} \cap \ considerados_{relevantes}\right)}{\#considerados_{relevantes}}\ , \qquad (2.42)$$

$$Recall = \frac{\#\left(termos_{relevantes} \cap considerados_{relevantes}\right)}{\#termos_{relevantes}}, \qquad (2.43)$$

onde, $termos_{relevantes}$ é o conjunto de termos verdadeiramente relevantes; $considerados_{relevantes}$ é o conjunto dos termos considerados relevantes pelo ordenador por grau de importância de prefixos, palavras e multipalavras no trabalho que construi. A quantidade de termos considerados relevantes pelo extractor e que são ao mesmo tempo realmente relevantes é representada por $\#\left(termos_{relevantes} \cap considerados_{relevantes}\right)$.

A Precisão (*Precision)* pode ser vista como medida de exactidão de uma ferramenta. Permite medir a proporção do número de termos realmente relevantes, dentro do conjunto dos termos que o extractor considera relevantes. Já a cobertura (*Recall)* mede a proporção do número de termos que, considerando o conjunto completo dos termos realmente relevantes, que foram detectados pelo extractor como tal.

Logo, no caso da avaliação dos resultados a serem gerados pela metodologia que se apresenta neste plano de trabalho são necessárias a *Precision* e o *Recall*, porque será necessário avaliar a correcção e completude dos resultados obtidos. É conveniente dizer que a avaliação da cobertura (*recall*) trará alguns problemas pois, à partida para os textos de onde irão ser extraídos termos chave, não existe um "*golden standard*" para nos informar da totalidade dos termos relevantes. No entanto, ao trabalhar e avaliar 6 medidas conseguimos obter um número de termos realmente relevantes, maior do que o número de termos relevantes que obteríamos se analisássemos apenas um método. De qualquer forma este número é inferior ao número total de termos relevantes pelo que calcularemos uma aproximação inferior ao *recall* real. Mas, de facto, é impossível olhar para todos os temos e classificá-los a todos como sendo relevantes ou não.

2.8.2 F-Measure

Esta medida é a média harmónica entre a Precision e Recall (ver secção 2.8.1), e é definida pela seguinte expressão

$$F - Measure = \frac{2 * precision * recall}{precision + recall} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}, \qquad (2{,}44)$$

O que combina desta forma os valores obtidos para a precision e para o recall.

A F-Measure apresenta valores elevados quando a *precision* e o *recall* apresentam valores elevados. Porque os valores de *recall* que obtemos são superiores ao verdadeiro *recall*, os valores de *F-Measure* que apresentarei são superiores ao verdadeiro *F-Measure.*

2.8.3 Estatística Kappa

A estatística Kappa (k) é uma medida estatística muito utilizada para avaliar o grau de concordância entre avaliações.

A matriz de confusão é um instrumento fundamental na análise e obtenção do valor da estatística kappa. Trata-se de uma matriz quadrada de dimensão NxN, em que N é o número de avaliações possíveis para um determinado termo.
No trabalho o resultado dessa avaliações podem ser cinco,

- GD – Good Descriptor
- NGD – Near Good Descriptor
- BD – Bad Descriptor
- U – Unkown
- NE – No Evalution

Na seguinte tabela, podemos ver uma representação de uma matriz de confusão para dois avaliadores, sobre um dado documento.

		Avaliador 1					
		GD	NGD	BD	U	NE	Total linha
Avaliador 2	GD	2	0	0	1	0	3
	NGD	0	0	0	0	0	0
	BD	0	0	1	0	0	1
	U	0	0	1	1	0	2
	NE	0	0	0	0	0	0
	Total Col	2	0	2	2	0	6

Tabela 2.2 – MCRV - Matriz Confusão com resultados verificados entre dois avaliadores

Onde na diagonal principal podemos encontrar o número de avaliações comuns entre os dois avaliadores para aquele documento. Por cada linha, por exemplo para a primeira linha, deve-se fazer a seguinte leitura:

- Posição [1, 1] – Número de termos avaliados como Good Descriptors por ambos os avaliadores;
- Posição [1, 2] – Número de termos avaliados como Good Decriptor pelo avaliador 2, mas como Near Good Desciptor pelo avaliador 1;

- Posição [1, 3] – Número de termos avaliados como Good Decriptor pelo avaliador 2, mas como Bad Desciptor pelo avaliador 1;
- Posição [1, 4] – Número de termos avaliados como Good Decriptor pelo avaliador 2, mas como Unknown pelo avaliador 1;
- Posição [1, 5] – Número de termos avaliados como Good Decriptor pelo avaliador 2, mas como No Evaluation pelo avaliador 1;

Para as restantes linhas e colunas, deve-se fazer leitura idêntica. Sendo que no caso das colunas, deve-se fazer a leitura para o avaliador 1 em função do avaliador 2.
Tendo obtido a matriz de confusão para os resultados verificados, é necessário calcular a matriz de confusão para os resultados esperados.
Esta matriz é preenchida tendo por base os valores da primeira matriz, onde cada posição desta nova matriz é preenchida pelo resultado da seguinte expressão:

$$MCRE_{Pos(i,j)} = \frac{\left(\sum Linha_i MCRV * \sum Coluna_j MCRV\right)}{\sum_{i=0}^{n} \sum Linha_i MCRV}, \quad (2.45)$$

		Avaliador 1					
		GD	NGD	BD	U	NE	Total linha
Avaliador 2	GD	1.2	0	0.6	1.2	0	3
	NGD	0	0	0	0	0	0
	BD	0	0	0	0	0	1
	U	0.8	0	0.4	0.8	0	2
	NE	0	0	0	0	0	0
	Total Col	2	0	2	2	0	6

Tabela 2.3 - MCRE Matriz Confusão com resultados esperados entre dois avaliadores

Exemplificando,

$$MCRE_{[1,1]} = \frac{(3 * 2)}{5} = 1.2$$

Tendo as duas matrizes de confusão calculadas, podemos então calcular a estatística kappa através da equação,

$$k = \frac{P_r(a) - P_r(e)}{trials - P_r(e)}, \quad (2.46)$$

No caso do trabalho realizado para a elaboração desta dissertação, $P_r(a)$, representa o somatório da diagonal principal da matriz de confusão dos resultados verificados pelos dois avaliadores. $P_r(e)$ Representa o somatório da diagonal principal da matriz de confusão para os valores esperados entres os dois avaliadores, e onde $trials$ é o numero total de termos avaliados pelos avaliadores.

Tendo o valor *kappa* calculado, vamos consultar a seguinte tabela de forma a identificar o grau de concordância entre os dois avaliadores.

Valor de Kappa	**Concordância**
< 0	Não existe concordância
0 – 0.20	Ligeira
0.21 – 040	Considerável
0.41 – 0.60	Moderada
0.61 – 0.80	Substancial
0.81 – 1	Excelente

Tabela 2.4 – Valores de K com a medida Estatística Kappa

2.9 Suffix Arrays

Text Mining a partir de texto não estruturado requer o uso de grandes quantidades de texto e o uso de estruturas suficientemente poderosas para a determinação das frequências de qualquer cadeia de caracteres, para indexação de textos completos, para reconhecimento de padrões e para extracção eficiente de cadeias de caracteres.

Suffix arrays[47], introduzida inicialmente como uma técnica de indexação de base de dados, é uma estrutura que tem sido bastante estudada ao longo das duas últimas décadas, capaz de suportar os requisitos acima descritos, visto que facilita a computação do cálculo da frequência e da localização de qualquer sub-cadeia de caracteres (um n-grama de caracteres, de palavras e de multipalavras) numa sequência longa de texto (corpus). Yamamoto e Church [8] estão entre vários autores que utilizam esta estrutura para a determinação de frequências de termos e de documentos para todos os n-gramas de dois grandes repositórios de texto. Seguidamente fazem uso destas frequências para calcular a Informação Mútua (*Mutual Information (MI))* entre

palavras para extraírem bigramas de palavras altamente coesos, candidatos a serem ou não multipalavras.

Uma das vantagens das Suffix Arrays relativamente às Suffix Trees é o espaço necessário. A necessidade de espaço por parte das Suffix Trees cresce com o tamanho do alfabeto: O (N |Σ|), onde |Σ| é o tamanho do alfabeto, ao contrário das Suffix Arrays, apesar de, em alfabetos de dimensão menor do que 24 caracteres, este factor ser pouco problemático. Manber e Myers [47] no seu trabalho afirmam que as suffix arrays estão numa ordem de magnitude mais eficiente no que diz respeito ao espaço ocupado em relação às suffx trees, mesmo no caso de alfabetos relativamente pequenos (|Σ| = 96). No entanto, nos últimos anos tem havido trabalhos nesta área que têm diminuído esta diferença entre estas duas estruturas, nomeadamente nos trabalhos [48] e [49].
Mas uma das motivações que me leva a se optar pelas Suffix arrays apesar destes avanços é o facto de poucos trabalhos nestas áreas fazerem uso desta estrutura para a extracção de tópicos ou palavras-chave relevantes.

Um Vector de sufixos, s, é um array de todos os N Sufixos ordenados alfabeticamente de um texto ou concatenação de textos. Um sufixo, s[i], também denominado por cadeia semi-infinita, é uma cadeia que começa no posição i do texto que estamos a tratar e continua até ao fim do mesmo.
A Figura 2.15 e a Figura 2.16 ilustram um exemplo simples, baseado no trabalho [8], onde o texto (“to_be_or_not_to_be”) é constituído por 18 sufixos (N = 18), 13 caracteres alfabéticos e 5 espaços, terminando a sequência com um terminador null. A Figura 2.15 mostra a inicialização do vector de sufixos. Já na Figura 2.16 vemos aquilo a que propriamente se chama suffix array ordenada. Porque os sufixos estão ordenados alfabeticamente

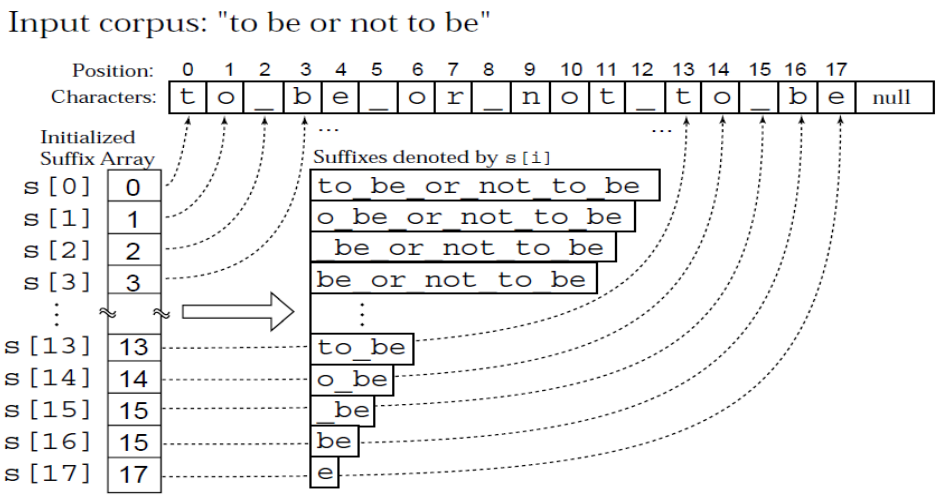


Figura 2.15 - Ilustração de uma Suffix Array, s, que acabou de ser inicializada e ainda não foi ordenada

Cada elemento da suffix array, s[i], é um inteiro que denota um sufixo ou uma string semi-infinita, a começar no posição i no texto até ao fim do texto. Exemplo baseado em[8]**.**

Suffix Array		Suffixes denoted by s[i]
s[0]	15	_be
s[1]	2	_be_or_not_to_be
s[2]	8	_not_to_be
s[3]	5	_or_not_to_be
s[4]	12	_to_be
s[5]	16	be
s[6]	3	be_or_not_to_be
s[7]	17	e
s[8]	4	e_or_not_to_be
s[9]	9	not_to_be
s[10]	14	o_be
s[11]	1	o_be_or_not_to_be
s[12]	6	or_not_to_be
s[13]	10	ot_to_be
s[14]	7	r_not_to_be
s[15]	11	t_to_be
s[16]	13	to_be
s[17]	0	to_be_or_not_to_be

Figura 2.16 - Ilustração da suffix array da **Figura 2.15** após ter sido ordenada.

Os inteiros em s são ordenados por forma a que as strings estejam alfabeticamente ordenadas. Exemplo baseado em[8]**.**

Como já foi dito anteriormente, as suffix arrays foram desenhadas para facilitar a computação e o cálculo das frequências de termos (*tf*) e apontar a localização de uma sub-string (ngrama/termo) numa sequência (texto). Dada uma sub-string ou termo, *t*, uma pesquisa binária é efectuada para encontrar o primeiro e o ultimo sufixo que começa com *t.* Seja s[i] o primeiro desses sufixos e s[j] o último. Então a frequência *tf*(*t*) = j – i + 1 e o termo está localizado nas posições : s do texto indicado.

A Figura 2.16 também mostra como é que este procedimento pode ser usados para calcular a frequência e para encontrar a localização de termos no corpus, veja-se o

exemplo de “to_be” no texto “to_be_or_not_to_be”. Como ilustrado também na Figura 2.16, s[i = 16] é o primeiro sufixo que começa com o termo "to_be" e s[j = 17] o último sufixo a começar com este termo.

Consequentemente, tf("to_be") = 17 - 16 + 1 = 2. Além disso as posições do termo "to_be" pode ser descrito como, posições ("to_be") = s = {13, 0}, e apenas estas posições.

Outra característica das suffix arrays é a de permitir encontrar o Prefixo Comum mais longo (*LCP*). Ou seja, permite a construção de um vector auxiliar de N + 1 inteiros. Em que cada lcp[i] indica o comprimento do prefixo comum entre s[i - 1] e s[i]. A Figura 2.17 exemplifica o vector dos lcp’s para a suffix array do texto “to_be_or_not_to_be”. O facto de lcp[11] ser igual a 4, significa que os prefixos de tamanho menor ou igual a 4 dos sufixos “o_be” ou “o_be_or_not_to_be” têm todos frequência 2 ou maior do que 2 (como acontece com o prefixo “o” que tem frequência 4). Qualquer prefixo de tamanho maior do que 4 de qualquer daqueles sufixos tem frequência 1.

Manber e Myers [47] fazem uso do vector de lcp’s para fazer a computação da frequência e encontrar a localização de uma sub-string de comprimento P numa sequência de comprimento N.

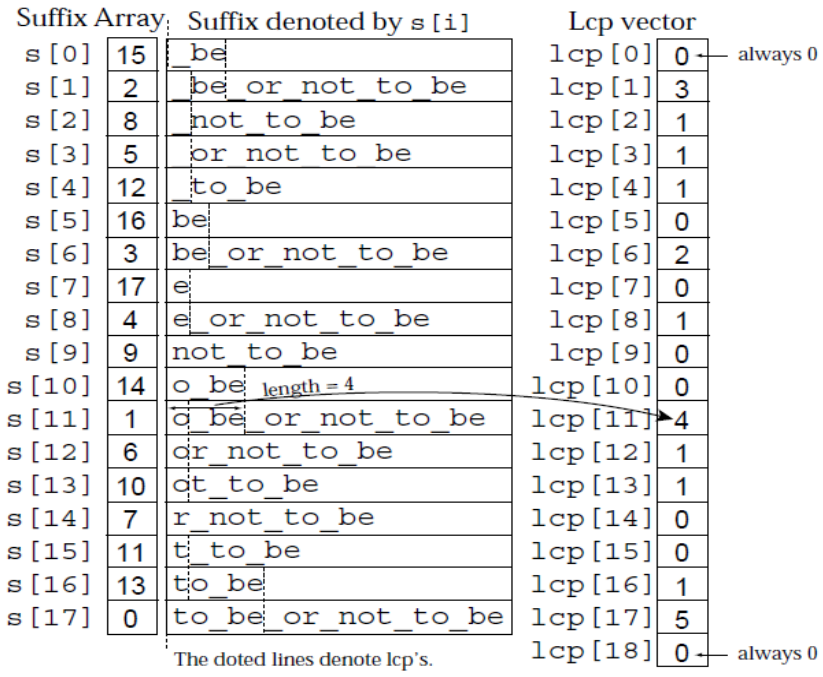


Figura 2.17 - O Prefixo comum mais longo (LCP)

O Prefixo comum mais longo (LCP) é um vector de N + 1 inteiros. *lcp*[i] denota o comprimento do prefixo comum entre o sufixo s[i – 1] e o sufixo s[i]. Por exemplo, s[10] e s[11] partilham um prefixo comum de 4 caracteres, portanto *lcp*[11] =

4. Nesta figura o prefixo comum está destacado a tracejado na suffix array e que é a mesma apresentada na **Figura 2.16**. Exemplo baseado em [8]

De acordo com Stefan Burkhardt e Juha Karkkainen[50] a construção de suffix arrays podem ser divididas em quatro categorias e segundo os mesmos os algoritmos de construção de Suffix arrays baseados em ordenação dos sufixos como strings independentes, como no exemplo apresentado, são a melhor opção para lidar com o problema que esta proposta de trabalho aborda. De facto, terei de determinar frequências de multipalavras, de palavras e de prefixos de 4 ou 5 caracteres de palavras, e respectiva localização poder aplicar qualquer das métricas de valorização dessas unidades textuais e, ao utilizar as Suffix arrays, bastar-me-á, percorrer a suffix array do início ao fim para ter imediatamente as características de que necessito.

Capitulo 3

Contribuição e Trabalho Realizado

3.1 Corpus de Teste

O Corpus de teste utilizado para a realização deste trabalho é composto por um conjunto de textos, em português, inglês e checo retirados da legislação europeia em vigor (http://eur-lex.europa.eu/pt/index.htm). Estes textos são os mesmos para as três línguas, com a ressalva de o checo ter mais nove documentos do que as outras duas línguas.

O primeiro passo, foi passar os textos de html para txt em UTF-8, esta tarefa foi realizada com recurso a um comando em Linux[42], como se indica a seguir:

"$> html2text -width 90 cs_32005D0754.html > cs_32005D0754.txt"

Onde html2text[43] é um comando onde especificamos o comprimento que as linhas do ficheiro de saída tinham de ter no máximo "*-width 90"* seguidamente especifica-se o ficheiro de entrada com a indicação do ficheiro de saída. *"cs_32005D0754.html > cs_32005D0754.txt".*
Esta tarefa foi realizada para todos os documentos do corpus.

[42] Distribuição Ubuntu 9.10
[43] http://manpages.ubuntu.com/manpages/intrepid/man1/html2text.1.html

A Dimensão do corpus em termos totais de termos, para cada língua que foi estudada é a seguinte:

Língua	Número de Termos	Número de Documentos
Português	109449	28
Inglês	100890	28
Checo	120787	37

Tabela 3.1 – Número de total de termos por Língua

3.2 Novas Medidas

Nesta secção apresentam-se as contribuições realizadas com esta dissertação, nomeadamente apresentando todas as variantes de medidas elaboradas no decorrer deste trabalho. Abordam-se primeiramente as versões das medidas *Tf-Idf*, *Phi-Square*, *Rvar* e Informação Mútua modificadas pelo operador *Least*, seguidamente as versões dessas medidas modificadas pelo operador *Bubbled*, foi introduzido também o operador Mediana, e no final, foram feitas combinações entre estes operadores.

3.2.1 Operador Least

As versões *Least* surgiram, pela necessidade de encontrar uma forma de ser possível comparar os resultados obtidos por J.F. da Silva no trabalho [1], para a medida LeastRvar (ver secção 2.3.1.2).

Assim, definimos que *Least* de uma medida para uma palavra seria o valor dessa medida para a própria palavra. Isto justifica-se porque o operador *Least* determinava o mínimo da medida *Rvar* para as duas palavras extremas de uma multipalavra. Para palavras resolvemos tratá-las como uma multipalavra em que a palavra é igual à palavra mais á esquerda desta pseudo multipalavra e é igual à palavra mais a direita dessa pseudo multipalavra.

Já quando tratamos de multipalavras, o valor *Least* será sempre o menor valor das medidas consideradas para as palavras nos extremos da multipalavra.

3.2.1.1 Least Tf-Idf

Seja W uma palavra,

Então,

$$Least\_TfIdf(W) = TfIdf(W) \qquad (3.1)$$

Se $W = w_1 \dots w_n$ for uma multipalavra.

Então,

$$Least\_TfIdf(w_1 \dots w_n) = Min\big(TfIdf(w_1), TfIdf(w_n)\big) \qquad (3.2)$$

Onde Min denota a função mínimo.

3.2.1.2 Least Rvar

Seja W uma palavra,

Então,

$$Least\_Rvar(W) = Rvar(W) \qquad (3.3)$$

Se $W = w_1 \dots w_n$ for uma multipalavra.

Então,

$$Least\_Rvar(w_1 \dots w_n) = Min\big(Rvar(w_1), Rvar(w_n)\big) \qquad (3.4)$$

Onde Min denota a função mínimo.

3.2.1.3 Least Chi Square

Seja W uma palavra,

Então,

$$Least\_ChiSquare(W) = ChiSquare(W) \qquad (3.5)$$

Se $W = w_1 \dots w_n$ for uma multipalavra.

Então,

$$Least\_ChiSquare(w_1 \dots w_n) = Min\big(ChiSquare(w_1), ChiSquare(w_n)\big) \qquad (3.6)$$

Onde Min denota a função mínimo.

3.2.1.4 Least Phi Square

Seja W uma palavra,

Então,

$$Least\_PhiSquare(W) \;=\; PhiSquare(W) \qquad (3.7)$$

Se $W = w_1 \dots w_n$ for uma multipalavra.

Então,

$$Least\_PhiSquare(w_1 \dots w_n) = Min\big(PhiSquare(w_1), PhiSquare(w_n)\big) \qquad (3.8)$$

Onde Min denota a função mínimo.

3.2.1.5 Least Informação Mútua (MI)

Seja W uma palavra,

Então,

$$Least\_MI(W) \;=\; MI(W) \qquad (3.9)$$

Se $W = w_1 \dots w_n$ for uma multipalavra.

Então,

$$Least\_MI(w_1 \dots w_n) \;=\; Min\big(MI(w_1), MI(w_n)\big) \qquad (3.10)$$

Onde Min denota a função mínimo.

3.2.2 Operador Bubbled

As versões *Bubbled* surgiram da necessidade de se associar um prefixo às palavras que sejam prefixadas por esse prefixo.

Esta variante só é aplicada directamente entre prefixos e palavras, não sendo feita a propagação a multipalavras. Esta propagação é efectuada aquando do uso de uma das seguintes variantes, *Least Bubbled* (secção 3.2.3) e *Least Bubbled Median* (secção 3.2.5).

Assim, o que foi feito foi o de associar a uma palavra, o valor da medida do prefixo que inicia essa palavra.

Por exemplo, suponhamos que o prefixo “multi” tem um valor para uma dada medida de 0.67.

E temos as palavras multilinguismo com o valor de 0.45, e a palavra multicultural com o valor de 0.78.

Como consequência do processo de *"Bubbling"* o valor de multilinguismo seria igual ao de multicultural que seria o valor do prefixo "multi" 0.67.

3.2.2.1 Bubbled TfIdf

Seja W uma palavra, e P ou um prefixo dessa palavra.

Então,

$$Bubbled\_TfIdf(W) = TfIdf(P) \qquad (3.11)$$

3.2.2.2 Bubbled Rvar

Seja W uma palavra, e P ou um prefixo.

Então,

$$Bubbled\_Rvar(W) = Rvar(P) \qquad (3.12)$$

3.2.2.3 Bubbled Chi Square

Seja W uma palavra, e P ou um prefixo dessa palavra.

Então,

$$Bubbled\_ChiSquare(W) = ChiSquare(P) \qquad (3.13)$$

3.2.2.4 Bubbled Phi Square

Seja W uma palavra, e P ou um prefixo dessa palavra.

Então,

$$Bubbled\_PhiSquare(W) = PhiSquare(P) \qquad (3.14)$$

3.2.2.5 Bubbled Informação Mútua

Seja W uma palavra, e P ou um prefixo dessa palavra.

Então,

$$Bubbled\_MI(W) = MI(P) \qquad (3.15)$$

3.2.3 Medidas Least Bubbled

Esta variante de medidas, partiu da necessidade de propagar as medidas *Bubbled* a multipalavras. Assim, esta variante passa pela combinação de fazer primeiro o Bubbling dos prefixos às palavras, e aplicar, depois a definição de *Least* a estes valores Bubbled.

3.2.3.1 Least Bubbled TfIdf

Seja W uma palavra, e P ou um prefixo dessa palavra.

Recorrendo a (3.11) obtemos,

$$Bubbled\_TfIdf(W) \;=\; TfIdf(P)$$

E aplicando a definição presente em (3.2), seja $W = w_1 \dots w_n$ uma multipalavra.

Então,

$$Least\ Bubbled\ TfIdf(W) = Min\big(Bubbled\_TfIdf(w_1), Bubbled\_TfIdf(w_n)\big) \qquad (3.16)$$

3.2.3.2 Least Bubbled Rvar

Seja W uma palavra, e P ou um prefixo dessa palavra.

Recorrendo a (3.12) obtemos,

$$Bubbled\_Rvar(W) \;=\; Rvar(P)$$

E aplicando a definição de LeastRvar (ver secção 2.3.1.2), seja $W = w_1 \dots w_n$ uma multipalavra.

Então,

$$Least\ Bubbled\ Rvar(W) = Min\big(Bubbled\_Rvar(w_1), Bubbled\_Rvar(w_n)\big) \qquad (3.17)$$

3.2.3.3 Least Bubbled Chi Square

Seja W uma palavra, e P ou um prefixo dessa palavra.

Recorrendo a (3.13) obtemos,

$$Bubbled\_ChiSquare(W) \;=\; ChiSquare(P)$$

E aplicando a definição presente em (3.3), seja $W = w_1 \dots w_n$ uma multipalavra.

Então,

$$Least\ Bubbled\ ChiSquare(W) = Min\big(Bubbled\_ChiSquare(w_1), Bubbled\_ChiSquare(w_n)\big) \qquad (3.18)$$

3.2.3.4 Least Bubbled Phi Square

Seja W uma palavra, e P ou um prefixo dessa palavra.

Recorrendo a (3.14) obtemos,

$$Bubbled\_PhiSquare(W) \;=\; PhiSquare(P)$$

E aplicando a definição presente em (3.6), seja $W = w_1 \dots w_n$ uma multipalavra.

Então,

$$Least\ Bubbled\ PhiSquare(W)$$
$$= Min\big(Bubbled\_PhiSquare(w_1), Bubbled\_PhiSquare(w_n)\big) \qquad (3.19)$$

3.2.3.5 Least Bubbled Informação Mútua

Seja W uma palavra, e P ou um prefixo dessa palavra.

Recorrendo a (3.15) obtemos,

$$Bubbled\_MI(W) = MI(P)$$

E aplicando a definição presente em (3.8), seja $W = w_1 \dots w_n$ uma multipalavra.

Então,

$$Least\ Bubbled\ MI(W)$$
$$= Min\big(Bubbled\_MI(w_1), Bubbled\_MI(w_n)\big) \qquad (3.20)$$

3.2.4 Medidas Least Median

Esta variante foi pensada para fazer uma comparação com a ideia expressa por J.F Silva em [7], que a Mediana de expressões relevantes faz com que expressões com maior mediana sejam melhor pontuadas.

A ideia que guia esta medida é a de aplicar a definição da operação *Least* (secção 3.2.1), e depois multiplicar este valor pela mediana do termo em questão, seja esse termo uma palavra, um prefixo ou uma multipalavra.

No que concerne ao cálculo da mediana, no caso de palavras, calculou-se este valor como sendo o comprimento da palavra.

No caso de se tratarem de multipalavras, temos de ter em conta o número de elementos da multipalavra a tratar e o tamanho desses elementos. Ou seja, tomemos como exemplo a seguinte multipalavra,

"Câmara Municipal de Murça"

É composta por 4 elementos, e o vector de tamanhos dos elementos resultante da multipalavra é:

V = {6,9,2,5}

Seguidamente ordenamos este vector, obtendo

Vord = {2,5,6,9}

Neste caso a mediana é dada pela seguinte operação,

Mediana = (5 + 6) / 2 = 5.5 = 6

No caso de se tratar de uma multipalavra, com um número impar de elementos, a operação altera-se, e efectua-se da seguinte forma,

"Assembleia da República"

É composta por 3 elementos, e o vector de tamanhos dos elementos resultante da multipalavra é:

V = {10,2,9}

Seguidamente ordenamos este vector, obtendo

Vord = {2,9,10}

Neste caso a mediana é dada pela seguinte operação,

Mediana = 9

Resumidamente, a mediana é dada pela seguinte expressão,

$$Mediana = \begin{cases} Compr.do\ termo, & se\ palavra. \\ \begin{cases} \dfrac{\left(Elem_{pos_{\left(\frac{n}{2}\right)}} + Elem_{pos_{\left(\frac{n}{2}\right)+1}}\right)}{2}, & se\ n\ par \\ Elem_{pos_{\left(\frac{n}{2}\right)}}, & se\ n\ impar \end{cases} & se\ multipalavra \end{cases} \quad (3.21)$$

Seja $Elem_{pos}$ um elemento pertencente ao Vector ordenado do tamanho das palavras de uma multipalavra. Com n compreendido entre o valor de um e o número de palavras da multipalavra.

3.2.4.1 Least Median TfIdf

Recorrendo à equação (3.1) ou (3.2) e aplicando um produto com a mediada, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Median\ TfIdf = Least\ TfIdf\ (W) * Median(W) \quad (3.22)$$

3.2.4.2 Least Median Rvar

Recorrendo à definição de *Rvar* e de *LeastRvar* (ver secção 2.3.1.2) e às equações (3.3) ou (3.4), aplicando um produto com a mediada, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Median\ Rvar = Least\ Rvar\ (W) * Median(W) \qquad (3.23)$$

3.2.4.3 Least Median Chi Square

Recorrendo á equação (3.5) ou (3.6) e aplicando um produto com a mediada, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Median\ ChiSquare = Least\ ChiSquare\ (W) * Median(W) \qquad (3.24)$$

3.2.4.4 Least Median Phi Square

Recorrendo á equação (3.7) ou (3.8) e aplicando um produto com a mediada, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Median\ PhiSquare = Least\ PhiSquare\ (W) * Median(W) \qquad (3.25)$$

3.2.4.5 Least Median Informação Mútua

Recorrendo á equação (3.9) ou (3.10) e aplicando um produto com a mediada, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Median\ MI = Least\ MI\ (W) * Median(W) \qquad (3.26)$$

3.2.5 Medidas Least Bubbled Median

No que concerne a esta variante de medida, pretendi verificar qual seria o impacto da mediana tendo o cálculo da medida LeastBubbled (ver secção 3.2.3) disponível.

Assim, esta medida é calculada obtendo o valor Least bubbled de um determinado termo, fazendo posteriormente o produto pela mediana do termo.

Seguem-se seguidamente a especificações para cada medida desta variante.

3.2.5.1 Least Bubbled Median TfIdf

Recorrendo á equação (3.16) e aplicando um produto com a mediana, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Bubbled\ Median\ TfIdf = Least\ Bubbled\ TfIdf\ (W) * Median(W) \qquad (3.27)$$

3.2.5.2 Least Bubbled Median Rvar

Recorrendo á equação (3.17) e aplicando um produto com a mediana, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Bubbled\ Median\ Rvar = Least\ Bubbled\ Rvar\ (W) * Median(W) \qquad (3.28)$$

3.2.5.3 Least Bubbled Median Chi Square

Recorrendo á equação (3.18) e aplicando um produto com a mediana, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Bubbled\ Median\ ChiSquare = Least\ Bubbled\ ChiSquare(W) * Median(W) \qquad (3.29)$$

3.2.5.4 Least Bubbled Median Phi Square

Recorrendo á equação (3.19) e aplicando um produto com a mediana, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Bubbled\ Median\ PhiSquare = Least\ Bubbled\ PhiSquare(W) * Median(W) \quad (3.30)$$

3.2.5.5 Least Bubbled Median Informação Mútua

Recorrendo á equação (3.20) e aplicando um produto com a mediana, obtemos a seguinte definição:

Seja W uma palavra ou multipalavra

Então

$$Least\ Bubbled\ Median\ MI = Least\ Bubbled\ MI(W) * Median(W) \quad (3.31)$$

3.3 Desenvolvimento

Nesta secção, irei descrever o ambiente de desenvolvimento, as ferramentas utilizadas, os problemas encontrados, bem como a descrição de opções e assunções tomadas ao longo da realização deste trabalho.

3.3.1 Ambiente de Desenvolvimento

O ambiente de desenvolvimento foi elaborado sobre o Sistema operativo Ubuntu[44] distribuição 9.10. A linguagem de programação utilizada foi Java[45], na versão 1.6_b13.

Os IDE's utilizados para o desenvolvimento do protótipo foram numa primeira fase o eclipse IDE[46]. Esta parte do desenvolvimento baseou-se no desenho das classes necessárias para implementar a ligação com a classe das Suffix Arrays utilizadas, as medidas utilizadas e os outputs dos primeiros testes.

[44] http://www.ubuntu.com/
[45] http://www.java.com/en/
[46] http://www.eclipse.org/

Quando surgiu a necessidade de se trabalhar a criação de um interface gráfico foi utilizado outro IDE, nomeadamente o Netbeans[47].

Houve um processo de importação de workspace de Eclipse para netbeans mantendo a código fonte num só local, sem ser necessária a duplicação de workspaces.

3.3.1.1 Suffix Arrays

A estrutura utilizada foi construída utilizando uma ponte em JNI[48] que permite a ligação de um módulo em C retirado de [51] e que possibilitou a sua utilização neste trabalho.

Para criar esta ligação, foram efectuados os seguintes passos:

- Primeiro criar o ficheiro *".class"* da classe Java onde temos implementado as chamadas ao módulo em C.
- Seguidamente utilizar o comando javah[49] para criar o ficheiro "*hearder"*[50]
    - $> javah sufArray.SuffixArray
- Depois de implementando o ficheiro jni (ver anexo 6, secção 6.1), é necessário efectuar a compilação destes mesmos ficheiros de forma a criar uma biblioteca binária que pode ser invocada em tempo de execução pelo java. Neste ponto houve uma dificuldade em efectuar estes passos em Windows.
    - Em Linux
        - gcc -c -shared -fpic -I/usr/lib/jvm/java-6-sun-1.6.0.20/include -I/usr/lib/jvm/java-6-sun-1.6.0.15/include/linux sarray.c scode.c ssarray.c lcp.c qsufsort.c SuffixArray.c
        - gcc -shared -I/usr/lib/jvm/java-6-sun-1.6.0.20/include -I/usr/lib/jvm/java-6-sun-1.6.0.15/include/linux sarray.o scode.o ssarray.o lcp.o qsufsort.o SuffixArray.o -o libsarray.so

[47] http://netbeans.org/
[48] Java Native Interface
[49] http://download.oracle.com/docs/cd/E17476_01/javase/1.4.2/docs/tooldocs/windows/javah.html
[50] Ficheiro .h em C.

- Em Windows
  - gcc -c -shared -Wl, -I"C:/Program Files/Java/jdk1.6.0_20/include" -I"C:/Program Files/Java/jdk1.6.0_20/include/win32" sarray.c scode.c ssarray.c lcp.c qsufsort.c sufArray_SuffixArray.c

  - gcc -shared -Wl,--kill-at -I"C:/Program Files/Java/jdk1.6.0_20/include" -I"C:/Program Files/Java/jdk1.6.0_20/include/win32" sarray.o scode.o ssarray.o lcp.o qsufsort.o sufArray_SuffixArray.o -o libsarrayWinVersion.dll

Realça-se aqui a necessidade que houve em ter que se recorrer a uma ferramenta denominada por MinGW[51] que fornece comandos gcc para Windows. Isto foi necessário porque as bibliotecas ".so" e ".dll" utilizadas em linux e Windows diferem entre si.

Com estes passos concluídos, através da utilização do seguinte método, é possível carregar em tempo de execução a biblioteca ".so" em linux ou a biblioteca ".dll" em Windows.

```
/**
 * Loads the C Library to the Java enviornment.
 */
public static void loadLibrary()
{
  String osName = System.getProperties().getProperty("os.name");
  if (osName.contains("Linux"))
  {
      final String library =
"/home/luis/workspace/Tese/SuffixArrays/src/sufArray/libsarray.so";
      System.load(library);
  } else if (osName.contains("Windows") || osName.contains("windows"))
  {
      final String library =
"c:/home/luis/workspace/Tese/SuffixArrays/src/sufArray/libsarrayWinVersion.
dll";
      System.load(library);
  }
}
```

[51] http://www.mingw.org/

3.4 Extracção de Palavras e Prefixos

A extracção de palavras e de prefixos do corpus que foi efectuada neste trabalho foi realizada da seguinte forma.
Primeiro foram lidos todos os ficheiros do corpus para uma variável *String* java, onde os textos de cada documento são separados por uma sequência de caracteres *" _^$#$#$^_"* pensada para este efeito.
Depois é construída uma Suffix Array[52] para esta String, recorrendo ao módulo C.

Recorremos a esta Suffix Array, para extrair as palavras da seguinte forma.
Percorrermos a SuffixArray, e só estamos interessados nas posições da suffixArray, cujo sufixo comece por um espaço em branco, esta condição indica-nos que o espaço em branco antecede sempre uma palavra, e consequentemente um prefixo.

Para ambas as situações uma segunda condição é avaliada, se a posição seguinte ao espaço em branco contem algum símbolo de pontuação, ou algum número. Se assim for não interessa, caso contrário, aplica-se um filtro que verifica se a palavra em questão tem um comprimento mínimo de seis caracteres. Se tiver seis ou mais caracteres, a palavra é considerada como válida e é inserida numa estrutura de dados desenhada para guardar a palavra com toda a informação associada a ela. No caso dos prefixos vamos verificar se no comprimento do prefixo candidato aparece algum espaço em branco, se aparecer não é prefixo e não interessa, caso contrário, o prefixo é inserido numa estrutura de dados desenhada para guardar o prefixo com toda a informação associada ao prefixo.

Estes métodos são apresentados, e têm como característica, a possibilidade de receber como parâmetro o comprimento mínimo que uma palavra deve ter e o número de caracteres que o prefixo deve ter, respectivamente. Ver no anexo 6 nas secções 6.2 e 6.3.

[52] Esta suffix array é retornada pelo módulo C já ordenada.

3.5 Extracção de Multipalavras

O Processo de extracção de multipalavras foi ligeiramente diferente das palavras e dos prefixos.

Foi aplicado um extractor[53] baseado em [2], sobre o texto tratado do corpus. Da seguinte linha de comandos resultou uma lista com as multipalavras (bigramas, trigramas, quadrigramas e pentagramas) de todo o corpus.

Esta lista é lida em tempo de execução e guardada na estrutura já mencionada na secção anterior, onde se insere uma palavra, neste caso multipalavra, com toda a informação associada a essa multipalavra.

$>cat ./Corpus/pt_txt/fixed_txt/.txt | ./relexp.py scp 5 | cut -f3 > MultiWordsList_ngrama_.txt*

Apesar de não fazermos uma extracção directa das multipalavras, aplicamos um filtro para que multipalavras que contenham números ou símbolos não sejam consideradas. Provavelmente, este tipo de filtro evita também que sejam avaliadas expressões desinteressantes como algumas que aparecem em [1] quando aqueles autores utilizaram a medida *Tf-Idf.*

Recorremos à mesma Suffix Array já apresentada na secção 3.4, com o objectivo de saber em que documentos as multipalavras aparecem e em que quantidade ocorrem nesses mesmo documentos. Com esta informação a multipalavra é inserida numa estrutura de dados desenhada para guardar esta a informação.

3.6 Implementação de Medidas

A implementação das medidas no protótipo desenvolvido tem duas partes distintas. A primeira é uma componente lógica, que recebendo todos os parâmetros necessários faz o cálculo da medida pretendida. Por exemplo, se quisermos calcular o valor do *Tf-Idf,* faríamos uso de uma classe estática Java, que recebe o valor do número de ocorrências do termo num determinado documento, o número total de termos nesse mesmo

[53] http://hlt.di.fct.unl.pt/luis/multiwords/index.html

documento, o número total de documentos e o número de documentos onde o termo a ser tratado ocorre, devolvendo posteriormente o valor para a medida.

Tendo este valor calculado, estamos prontos para usar a segunda parte que compõe a parte de implementação das medidas. Na estrutura desenvolvida, cada termo tem como membro privado da sua classe um objecto que representa uma determinada medida, onde vamos guardar os valores das medidas calculados, como o *Tf-Idf* acima descrito. Isto possibilita a persistência dos dados, em suporte físico, permitindo também que o cálculo das medidas seja feita uma só vez, na inicialização das estruturas, caso não existam em suporte físico, no arranque do protótipo.

3.7 Protótipo

Nesta secção, pretende-se dar uma visão mais global sobre o protótipo que foi idealizado e realizado no decorrer deste trabalho. Foi desenhado para permitir a uma interacção mais "*user friendly*" entre o avaliador e o texto base, e os descritores a classificar. Também foi implementado uma interface que permite analisar os resultados dos diferentes avaliadores e perceber quais os valores de Precisão, Cobertura e F-Measure (secções 2.8.1 e 2.8.2) associados a estes resultados. Possibilita ainda verificar a estatística Kappa (secção 2.8.3) entre dois avaliadores. O manual do utilizador do protótipo é apresentado no anexo 2, secção 7 desta dissertação.

3.7.1 Desenho e Diagrama do protótipo

Apresenta-se de seguida um diagrama de pacotes que ilustra o desenho adoptado na implementação do protótipo.

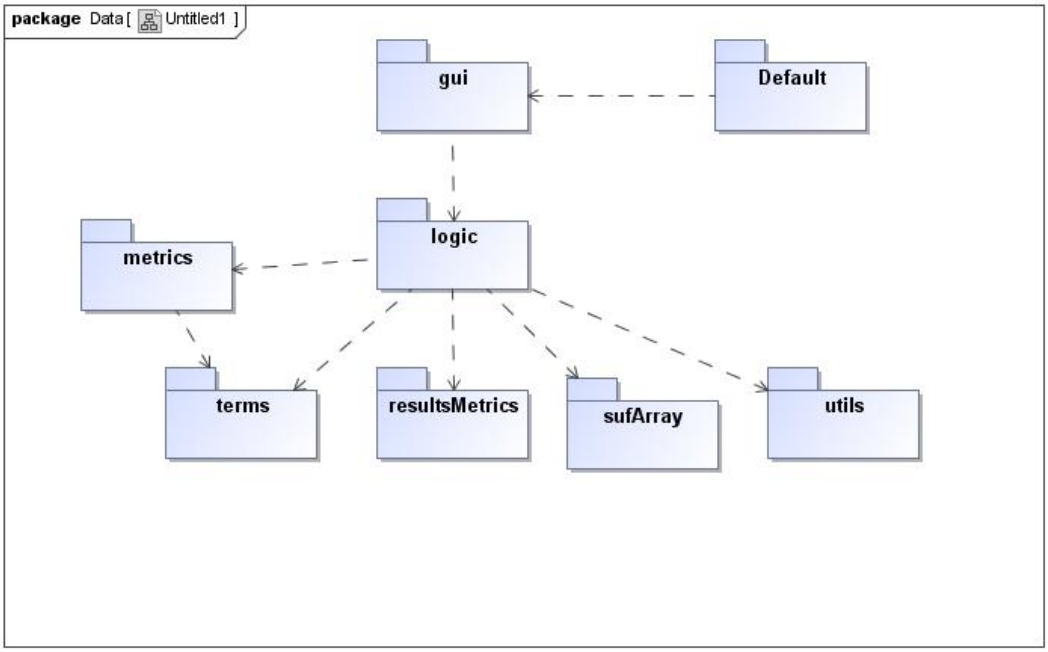


Figura 3.1 – Diagrama de Pacotes do Protótipo.

Como se pode verificar pela figura anterior, adaptou-se uma organização em três níveis no desenvolvimento do protótipo. Assim, qualquer desenvolvimento posterior será localizado num só pacote, e estanque nas repercussões pelo resto do código desenvolvido.

3.8 Considerações

3.8.1 Considerações sobre Trabalho Realizado

No trabalho que desenvolvi fiz uso de uma suffix array ordenada. Isto permitiu-me fazer a extracção de palavras e de prefixos de uma forma muito rápida e eficiente. Permitiu-me também usar a suffix array para encontrar onde as multipalavras extraídas pelo extractor utilizado ocorriam. Mais detalhes obvre possíveis melhoramentos podem ser encontrados no capítulo 5.

3.8.2 Considerações sobre Contribuições

Além das medidas base, tornou-se necessário a criação de outras medidas derivadas (secção 3.2) das medidas base. Estas novas medidas mostraram alguns resultados interessantes, como se poderá ver mais em pormenor no capítulo 4.

Capitulo 4

Resultados Obtidos e sua Avaliação

Neste capítulo apresentam-se alguns resultados e faz-se uma discussão dos mesmos. Faremos algumas considerações sobre as medidas base, discutindo algumas leituras que foram possível fazer ao longo da experimentação efectuada. Além disso, apresentar-se-ão também alguns resultados que se consideram interessantes do ponto de vista da experimentação.

Convém salientar que a cada avaliador foi pedido que avaliasse obrigatoriamente 25 termos para seis medidas distintas, nomeadamente, *Phi-Square, Least Tf-Idf, Least Median RVar, Least Median MI, Least Bubbled Median Phi-Square e Least Bubbled Median Rvar*. Estas são as primeiras seis *tabs* apresentadas na aplicação dos avaliadores (ver Figura 7.23 no Anexo 2, secção 7). Note-se que, ao passar de medida para medida, o avaliador já vai ter termos anteriormente avaliados, especialmente no que toca às medidas baseadas em *Tf-Idf* e *Phi-Square.* Observa-se que quando se passa para as medidas baseadas em *Rvar* ou *MI*, o número de termos não avaliados é considerável. Mas as variantes destas duas medidas acabam por dar resultados bastante semelhantes.

Convém dizer que a escolha destas medidas, para serem avaliadas pelos avaliadores, foi feita com base nos os resultados preliminares que se foram verificando ao longo do desenvolvimento do trabalho. Acresce ainda que tínhamos que limitar a quantidade de trabalho pedida aos avaliadores. Assim, e para que uma amostra representativa de todos os tipos de medidas utilizadas, escolheu-se o *Phi-Square* para o tipo de medida base, a medida *Least Tf-Idf* para uma medida com o operador *Least.* Para a conjugação

de operadores escolhemos *Least Median RVar, Least Median MI, Least Bubbled Median Phi-Square e Least Bubbled Median Rvar*.
A escolha da medida *Least Median Rvar* tinha de ser feita pois era a medida que havia sido considerada como a melhor em [7].
Não escolhemos uma medida só com o operador *"Bubble" porque o efeito "Bubbled"* pode ser verificado nas medidas *Least Bubbled Median* esolhidas.

Iremos tomar como exemplos, alguns ficheiros do corpus que foram avaliados, apresentaremos as listagens de termos que foram apresentados aos avaliadores, apresentaremos as avaliações que os avaliadores deram a esses mesmo termos, correlacionaremos o grau de concordância entre cada dois avaliadores que avaliaram o mesmo documento, através da apresentação do valor Kappa (ver secção 2.8.3).

Para cada língua utilizada na experimentação, vamos apresentar um documento em comum para dois avaliadores. O checo é uma excepção porque só conseguimos a avaliação por um único avaliador.

4.1 Língua Portuguesa

Começamos por apresentar resultados para a língua Portuguesa.

Apresentamos seguidamente as avaliações feitas pelos avaliadores Prof. Joaquim Ferreira da Silva e Prof. Gabriel Lopes. Um documento avaliado por ambos é o pt_32006R0198.html[54] Para as várias medidas, que foram pedidas para serem avaliadas obrigatoriamente, estes autores obtiveram os valores de precisão que são apresentados para as várias medidas ao longo das próximas secções.

4.1.1 Phi-Square

No caso do *Phi-Square*, a listagem de termos produzida, que foi apresentado aos avaliadores é a seguinte:

Termos	**Valor da medida**
formação profissional contínua	0,008977472052384
profissional contínua	0,008977472052384
contínua	0,008257084363260
formação profissional	0,007613838869853
profissional	0,006731434220435
em horas	0,005207533750025
cursos de formação profissional contínua	0,005096688636165
cursos	0,005080076295244
cursos de formação	0,005064663891633
formação	0,004140313788898
nenhum valor em falta	0,003545069493752
valor em falta	0,003545069493752
nenhum valor	0,003545069493752
número	0,003345129880868
número total	0,003309304724491
imputação	0,002547809415785
profissional inicial	0,002534794484038
tempo de trabalho	0,002437012852767
remunerado	0,002437012852767
nenhum	0,002421652204649
empresas	0,002204694848287
amostragem	0,002200631608461
inicial	0,002125444852977
empregadas	0,002120291214962
— sem classificação	0,001883060370466

Tabela 4.1 – Lista de Termos para a medida Phi-Square para o ficheiro pt_32006R0198.html

[54] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

Desta listagem podemos observar que a medida *Phi-Square* dá uma pontuação diferenciada a praticamente todos os termos. Em Anexo nas secções 8.2.1 e 8.3.1, podemos ver como os avaliadores avaliaram esta lista de termos.

No caso desta medida, as precisões obtidas foram as que se apresentam a seguir.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0,2	0,8	0.061224489795918	0.111111111111111
10	0,7	0,1	0,8	0.142857142857143	0.237288135593220
15	0,4666667	0,266666667	0,733333	0.142857142857143	0.218750000000000
20	0,45	0,2	0,65	0.183673469387755	0.260869565217391

Tabela 4.2- Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Phi-Square

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0	0,4	0,066666667	0,114285714
10	0,6	0	0,6	0,2	0,3
15	0,4	0	0,4	0,2	0,266666667
20	0,368421053	0	0,368421053	0,233333333	0,285714286

Tabela 4.3 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Phi-Square

Como podemos constatar da análise das tabelas Tabela 4.2 e Tabela 4.3, a precisão total, que tem em conta a precisão de bons descritores somada com a precisão dos quase bons descritores, obtida pelos avaliadores é bastante próxima. Apesar de as parcelas da soma serem distintas entre os dois.

Obtemos para esta medida um valor de Kappa de 0,552429667519181, valor que dá aproximadamente 55.2% de concordância, ou seja, uma concordância moderada de acordo com a Tabela 2.4 No Anexo 3, na secção 8.1.1 podemos ver as matrizes necessárias na obtenção deste valor.

Na secção 8.5 e 8.10, podemos ver os gráficos das precisões obtidas dos resultados destes avaliadores para o documento e medida em causa.

Podemos constatar pela Tabela 4.1, que de facto os termos extraídos por esta medida dão uma boa pista sobre o conteúdo do documento em causa. Veja-se por exemplo o termo mais bem classificado, "formação profissional contínua". Tendo em consideração a leitura do documento em causa verifica-mos que se trata de facto de um documento sobre formação profissional.

4.1.2 Least Tf-Idf

No caso desta medida, a listagem de termos que foi apresentado aos avaliadores é a seguinte:

Termos	**Valor da medida**
profissional	0,017270167990526
contínua	0,016727894319951
profissional contínua	0,016727894319951
cursos de formação profissional contínua	0,012184515615767
cursos	0,012184515615767
formação profissional contínua	0,009593030169595
formação	0,009593030169595
cursos de formação	0,009593030169595
formação profissional	0,009593030169595
cursos internos de formação	0,009593030169595
imputação	0,009187329625273
formação específicas das pessoas empregadas	0,009174378153781
contínua para pessoas empregadas	0,009174378153781
empregadas	0,009174378153781
empresas	0,008973854651220
empregadas em empresas	0,008973854651220
profissional nas empresas	0,008973854651220
formação profissional nas empresas	0,008973854651220
empresas que fazem formação	0,008973854651220
remunerado para cursos	0,008787880511131
remunerado	0,008787880511131
remunerado em cursos	0,008787880511131
participantes em cursos	0,006961567700693
participantes	0,006961567700693
participantes em formação profissional	0,006961567700693

Tabela 4.4 – Lista de Termos para a medida Least Tf-Idf para o ficheiro pt_32006R0198.html

Desta listagem podemos observar que a variante *Least Tf-Idf* apresenta uma certa dificuldade em diferenciar alguns termos, sendo que neste caso é fruto da definição de *Least Tf-Idf*. Podemos observar na Tabela 4.4, grupos de termos com a mesma pontuação. Apesar disso, é possível diferenciar uma certa hierarquização nos resultados. Em Anexo nas secções 8.2.2 e 8.3.2, podemos ver como os avaliadores avaliaram esta lista de termos.

No que concerne aos valores de precisão obtidos para esta medida, podemos ver as seguintes tabelas:

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0,2	0,8	0.062500000000000	0.113207547169811
10	0,8	0,1	0,9	0.166666666666667	0.275862068965517
15	0,733333333	0,133333333	0,866666667	0.229166666666667	0.349206349206349
20	0,6	0,2	0,8	0.250000000000000	0.352941176470588

Tabela 4.5- Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Tf-Idf

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0	0,4	0.068965517241379	0.117647058823529
10	0,7	0	0,7	0.241379310344828	0.358974358974359
15	0,5714286	0,071428571	0,642857143	0.275862068965517	0.372093023255814
20	0,5263158	0,052631579	0,578947368	0.344827586206897	0.416666666666667

Tabela 4.6 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Tf-Idf

Da análise destas tabelas, podemos ver um desvio nos valores de precisão para o resultado dos avaliadores. Enquanto que na Tabela 4.5 observamos uma precisão total muito boa, já o mesmo não se pode dizer comparativamente da Tabela 4.6, apesar dos valores de precisão para 10,15 e 20 passarem o valor de 0.5.

Para esta medida os autores tem um valor de concordância de 0.63235, o que dá aproximadamente 63.24%, o que de acordo com a tabela de concordância apresentada na secção 2.8.3, temos uma concordância Substancial. No Anexo 3, na secção 8.1.2 podemos ver as matrizes necessárias na obtenção deste valor.

4.1.3 Least Median Rvar

No caso desta medida, a listagem de termos que foi apresentado aos avaliadores é a seguinte:

Termos	Valores da medida
estatísticas-chave	17,999999999999996
significativamente	17,999999999999996
pormenorizadamente	17,999999999999996
subpopulações-alvo	17,999999999999996
electronicamente	15,999999999999996
horvitz-thompson	15,999999999999996
socioeconómicas	14,999999999999996
variáveis-chave	14,999999999999996
variável-chave	14,000000000000000
estratificados	13,999999999999996
probabilística	13,999999999999996
corresponderam	13,999999999999996
pormenorizados	13,999999999999996
população-alvo	13,999999999999996
sobrecobertura	13,999999999999996
significativamente melhorados	13,999999999999996
probabilística estratificada	13,499999999999996
variável-base	13,000000000000000
empresas-mães	12,999999999999996
laboratoriais	12,999999999999996
preenchimento	12,999999999999996
destacamentos	12,999999999999996
identificadas	12,999999999999996
não-respostas	12,999999999999996
problemáticas	12,999999999999996

Tabela 4.7 – Lista de Termos para a medida Least Median Rvar para o ficheiro pt_32006R0198.html

Desta listagem podemos observar que a variante *Least Median Rvar* apresenta uma maior dificuldade em hierarquizar termos. Podemos observar na Tabela 4.7, grupos de termos com a mesma pontuação. Apesar disso, é possível diferenciar uma hierarquização nos resultados ao contrário da medida base *Rvar*, que não possibilita diferenciação nenhuma, como veremos mais em pormenor no capítulo 5

Em Anexo nas secções 8.2.3 e 8.3.3, podemos ver como os avaliadores avaliaram esta lista de termos.

No que concerne aos valores de precisão obtidos para esta medida, podemos ver as seguintes tabelas:

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0	0,4	0.040816326530612	0.074074074074074
10	0,5	0,2	0,7	0.102040816326531	0.169491525423729
15	0,4666667	0,2	0,666666667	0.142857142857143	0.218750000000000
20	0,45	0,25	0,7	0.183673469387755	0.260869565217391

Tabela 4.8 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Median Rvar

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,2	0,4	0.033333333333333	0.057142857142857
10	0,4	0,1	0,5	0.133333333333333	0.200000000000000
15	0,2666667	0,266666667	0,533333333	0.133333333333333	0.177777777777778
20	0,25	0,2	0,45	0.166666666666667	0.200000000000000

Tabela 4.9 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median Rvar

Para esta medida os autores têm um valor de concordância de 0.10913, o que dá aproximadamente 11%, o que de acordo com a tabela de concordância apresentada na secção 2.8.3, temos uma concordância ligeira.

Este facto deve-se sobretudo à dictomia de critérios que pode ser observado na Tabela 8.15 e na Tabela 8.21, onde podemos verificar que existe uma maior consideração de "*Near good descriptors*" por parte do avaliador Prof. Joaquim da Silva Ferreira, que são considerados como *"Bad Descriptors"* por parte do avaliador Prof. Gabriel Lopes.

No Anexo 3, na secção 8.1.3 podemos ver as matrizes necessárias na obtenção deste valor.

4.1.4 Least Median MI

No caso desta medida, a listagem de termos que foi apresentado aos avaliadores é a seguinte:

Termos	Valores da Medida
estatísticas-chave	46,359290347154630
significativamente	46,359290347154630
pormenorizadamente	46,359290347154630
subpopulações-alvo	46,359290347154630
electronicamente	41,208258086359670
horvitz-thompson	41,208258086359670
socioeconómicas	38,632741955962190
variáveis-chave	38,632741955962190
estratificados	36,057225825564714
probabilística	36,057225825564714
corresponderam	36,057225825564714
pormenorizados	36,057225825564714
variável-chave	36,057225825564714
população-alvo	36,057225825564714
sobrecobertura	36,057225825564714
significativamente melhorados	36,057225825564714
probabilística estratificada	34,769467760365970
empresas-mães	33,481709695167230
laboratoriais	33,481709695167230
preenchimento	33,481709695167230
destacamentos	33,481709695167230
identificadas	33,481709695167230
não-respostas	33,481709695167230
problemáticas	33,481709695167230
questionários	33,481709695167230

Tabela 4.10 - Lista de Termos para a medida Least Median MI para o ficheiro pt_32006R0198.html

À semelhança da medida anterior, secção 4.1.3, esta medida também apresenta uma maior dificuldade em hierarquizar termos. Podemos observar na Tabela 4.8, grupos de termos com a mesma pontuação. Apesar disso, é possível diferenciar uma hierarquização nos resultados ao contrário da medida base *MI*, como veremos mais em pormenor no capítulo 5

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0	0,4	0.040816326530612	0.074074074074074
10	0,4	0,3	0,7	0.081632653061224	0.135593220338983
15	0,4666667	0,2	0,666666667	0.142857142857143	0.218750000000000
20	0,45	0,25	0,7	0.183673469387755	0.260869565217391

Tabela 4.11- Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Median MI

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,2	0,4	0.033333333333333	0.057142857142857
10	0,3	0,2	0,5	0.100000000000000	0.150000000000000
15	0,266666667	0,266666667	0,533333333	0.133333333333333	0.177777777777778
20	0,2	0,2	0,4	0.133333333333333	0.160000000000000

Tabela 4.12 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median MI

Para esta medida os autores tem um valor de concordância de 0,0196, o que dá aproximadamente 1.96%, o que de acordo com a tabela de concordância apresentada na secção 2.8.3, temos uma concordância ligeira.

4.1.5 Least Bubbled Median Phi-Square

No caso desta medida, a listagem de termos que foi apresentado aos avaliadores é a seguinte:

Termos	Valores da Medida
contínua	0,062639410875556
profissional	0,056544502411978
profissional contínua	0,047120418676649
formação profissional	0,041244206779647
empresas-mães	0,040936954447726
cursos de formação profissional contínua	0,040640610361951
amostragem	0,038514649217131
amostrais	0,034663184295418
empresarial	0,034638961455768
formação profissional contínua	0,032995365423718
formação	0,032995365423718
variáveis-chave	0,032924777689455
amostragem incluídas na amostra	0,030811719373705
variável-chave	0,030729792510158
cursos	0,030480457771463
cursos internos de formação	0,028870944745753
variável-base	0,028534807330861
formação no desempenho empresarial	0,028340968463810
imputações	0,027694086088190
amostra	0,026960254451992
empresas nos estratos de amostragem	0,025191971967832
empresas	0,025191971967832
profissional nas empresas	0,025191971967832
formação profissional nas empresas	0,025191971967832
formação profissional contínua da empresa	0,025191971967832

Tabela 4.13 - Lista de Termos para a medida Least Bubbled Median Phi-Square para o ficheiro pt_32006R0198.html

À semelhança do que já tinha acontecido com a medida *Phi-Square* esta variante apresenta também uma boa hierarquização de termos pelo valor da medida. Apesar de nas ultimas posições da Tabela 4.13 haver uma sequencia de 5 termos com o mesmo valor de medida.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0,2	0,8	0.062500000000000	0.113207547169811
10	0,6	0,3	0,9	0.125000000000000	0.206896551724138
15	0,7333333	0,2	0,933333333	0.229166666666667	0.349206349206349
20	0,8	0,15	0,95	0.333333333333333	0.470588235294118

Tabela 4.14 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Bubbled Median Phi-Square

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,0	0,2	0.034482758620690	0.058823529411765
10	0,5	0,0	0,5	0.172413793103448	0.256410256410256
15	0,6	0,0	0,6	0.310344827586207	0.409090909090909
20	0,684210526	0,0	0,684210526	0.448275862068966	0.541666666666667

Tabela 4.15 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Phi-Square

Para esta medida os autores tem um valor de concordância de 0,634502923976608, o que dá aproximadamente 63.45%, o que de acordo com a tabela de concordância apresentada na secção 2.8.3, temos uma concordância substancial.

4.1.6 Least Bubbled Median Rvar

No caso desta medida, a listagem de termos que foi apresentado aos avaliadores é a seguinte:

Termos	Valores da Medida
subpopulações-alvo	17,999999999999996
horvitz-thompson	15,999999999999996
não-respostas	13,000000000000004
destacamentos	12,999999999999996
influenciaram	12,999999999999996
não-resposta	12,000000000000004
reponderação	11,999999999999996
não-formação	11,999999999999996
pac=c3tot*a5	11,999999999999996
coeficientes	11,999999999999996
subcobertura	11,999999999999996
planificação	11,999999999999996
acessibilidade	11,943045311153242
comentários	11,000000000000002
coeficiente	10,999999999999998
codificação	10,999999999999998
sobrecobertura	10,842529794442926
probabilística	10,383412029287300
ventilação	10,000000000000002
honorários	10,000000000000002
calcula-se	10,000000000000000
imputações	10,000000000000000
calcularão	10,000000000000000
subamostra	9,999999999999998
recalcular	9,999999999999998

Tabela 4.16 - Lista de Termos para a medida Least Bubbled Median Rvar para o ficheiro pt_32006R0198.html

À semelhança do que aconteceu com a primeira variante do Rvar que vimos, na secção 4.1.3, a medida que estamos a analisar também apresenta dificuldades na hierarquização dos termos pelos valores obtidos na medida. Podemos ver dois grandes grupos na Tabela 4.16, uma grupo de 6 termos com o valor de 11.99 e um grupo de 5

termos com o valor de 10.0, isto faz com se veja dois grupos, sem uma clara hierarquização. Apesar de tudo, sempre apresenta resultados mais aceitáveis que a medida Rvar, como veremos em mais pormenor no capítulo 5.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,8	0	0,8	0.085106382978723	0.153846153846154
10	0,8	0	0,8	0.170212765957447	0.280701754385965
15	0,8666667	0	0,866666667	0.270833333333333	0.412698412698413
20	0,85	0,05	0,9	0.354166666666667	0.500000000000000

Tabela 4.17 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Bubbled Median Rvar

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,5	0,25	0,75	0.066666666666667	0.117647058823529
10	0,5	0,125	0,625	0.133333333333333	0.210526315789474
15	0,333333333	0,166666667	0,5	0.133333333333333	0.190476190476190
20	0,235294118	0,235294118	0,470588235	0.133333333333333	0.170212765957447

Tabela 4.18 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Rvar

Para esta medida os autores tem um valor de concordância de 0,2152466367713, o que dá aproximadamente 21.52%, o que de acordo com a tabela de concordância apresentada na secção 2.8.3, temos uma concordância considerável.

4.2 Leitura de Resultados para a Língua Portuguesa

Do que pudemos constatar pela leitura dos resultados obtidos da avaliação efectuada pelos avaliadores. Podemos destacar de imediato que três medidas apresentam um grau de concordância substancial, nomeadamente o *Phi-Square*, o *Least Tf-Idf* e o *Least Bubbled Median Phi-Square*.

Constata-se também, que estas mesmas medidas apresentam termos com maior significado semântico que as outras medidas avaliadas, nas quais predominam muito verbos, adjectivos e advérbios.

Mais, sabendo que o avaliador Prof. Gabriel Lopes avaliou uma amostra de nove documentos. As precisões totais médias, obtidas para as medidas que foram avaliadas na totalidade pode ser visto na seguinte Tabela 4.19. Onde podemos observar que em média, a precisão total mais elevada para todos os limites considerados (5,10,15,20) são obtidos pelas medidas *Phi-Square*, *Least-Tf-Idf* e *Least Bubbled Median Phi-Square*. Podemos ver uma ilustração da distribuição da precisão total pelos documentos avaliados pelo avaliador na secção 8.6.

Precision \ Threshold	**Phi^2**	**Least Tf-Ifd**	**Least M Rvar**	**Least M MI**	**Least M B Phi^2**	**Least M B Rvar**
T. Prec. Avg (5)	0,727777778	0,638888889	0,462962963	0,424074074	0,622222222	0,516666667
T. Prec. Avg (10)	0,725	0,660978836	0,355202822	0,353968254	0,613580247	0,483289242
T. Prec .Avg (15)	0,68026048	0,640761091	0,347985348	0,351628002	0,62049062	0,453106153
T. Prec. Avg (20)	0,621251386	0,645621202	0,345351328	0,334064942	0,626377422	0,414740896

Tabela 4.19 – Precisões Totais médias para Português para o Avaliador Prof. Gabriel Lopes

Na secção 8.7, podemos ver gráficos que apresentam a relação entre a precisão total de cada documento e a média da precisão. Estes gráficos só conseguem ser produzidos para um limite de cada vez, ou seja, para se observar o comportamento da precisão para os vários limites, seria necessário fazer quatro gráficos distintos. Devido a esse facto, optou-se por mostrar os gráficos para o limite 5 e 20. A amostra de gráficos não será exaustiva para todas as medidas, mas somente a algumas que apresentam melhores resultados de precisão e a algumas que apresentam piores resultados de precisão.
Uma outra leitura que podemos fazer, dos gráficos ilustrados da Figura 8.17 à Figura 8.20 é a de que a medida *Least Median Rvar* e a medida *Least Median MI* apresentam muitas semelhanças em termos da precisão dos documentos em relação à média.

Já na secção 8.8 podemos ver a média de precisão total para todas as medidas desenvolvidas nesta dissertação, pelos resultados das avaliações do avaliador Prof. Gabriel Lopes. Na qual podemos constatar que os resultados para as medidas base, *Rvar* e *MI*, bem como algumas variantes destas mesmas medidas (com excepção das que foram obrigatoriamente avaliadas) não apresentam resultados. Isto deve-se aos maus resultados produzidos por estas medidas. Como podemos ver na Tabela 8.25 e na Tabela 8.26 de termos apresentados aos avaliadores para a medida *Rvar* e *MI*, respectivamente., verificamos que não apresentam muitos termos em comum com as

suas variantes (Tabela 4.7, Tabela 4.10 e Tabela 4.16), daí a propagação de avaliações de possíveis termos comuns torna-se impraticável.
Outra leitura que podemos constar da Tabela 8.25 e da Tabela 8.26 é a incapacidade do *Rvar* e do *MI* de conseguirem fazer uma diferenciação de termos. Todos os termos tem o mesmo valor de medida, isto torna uma hierarquização de termos impossível pelo seu peso.

O avaliador Prof. Joaquim Ferreira da Silva avaliou uma amostra de cinco documentos. As precisões totais médias, obtidas para as medidas que foram avaliadas na totalidade pode ser visto na seguinte Tabela 4.20. Podemos observar também que em média, a precisão total mais elevada para todos os limites considerados (5,10,15,20) são obtidos pelas medidas *Phi-Square* e *Least Bubbled Median Phi-Square*.

Prec \ Threshold	**Phi^2**	**Least Tf-Ifd**	**Least M Rvar**	**Least M MI**	**Least M B Phi^2**	**Least M B Rvar**
T. Prec. Avg (5)	0,84	0,56	0,76	0,72	0,76	0,8
T. Prec. Avg (10)	0,8	0,7	0,72	0,74	0,66	0,66
T. Prec. Avg (15)	0,746666667	0,706666667	0,64	0,64	0,68	0,605714286
T. Prec. Avg (20)	0,75	0,73	0,62	0,63	0,68	0,614210526

Tabela 4.20 – Precisões Totais médias para Português para o Avaliador Prof. Joaquim Ferreira da Silva

Na secção 8.13 podemos ver a média de precisão total para todas as medidas desenvolvidas nesta dissertação, pelos resultados das avaliações do avaliador Prof. Joaquim Ferreira da Silva. Podemos constatar também que os resultados para as medidas base, *Rvar* e *MI*, bem como algumas variantes destas mesmas medidas, em menor quantidade que as do avaliador anterior e com excepção das que foram obrigatoriamente avaliadas, não apresentam resultados. Isto deve-se, como já foi dito, ao facto de as medidas base *Rvar* e *MI* não apresentarem muitos termos em comum com as suas variantes.

A diferenciação dos resultados entre estes dois autores, deve-se ao facto de que, por parte do avaliador Prof. Joaquim Ferreira da Silva o uso da classificação de *"Near Good Descriptor*" foi mais usado do que por parte do avaliador Prof. Gabriel Lopes. Este facto, pode ser constatado pelas tabelas das avaliações efectuadas pelos mesmos, no anexo 2, nas secções 8.2 e 8.3. Este facto leva a que as precisões totais médias

alcançadas para o avaliador Prof. Joaquim Ferreira da Silva sejam mais equitativas entre as medidas.

No que diz respeito à cobertura média alcançada por parte destes avaliadores, podemos ver as seguintes tabelas.

Recall \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0,162332188	0,140275652	0,057282204	0,061528327	0,136911887	0,055350608
Recall Avg (10)	0,303927597	0,245604161	0,079072186	0,078817157	0,234905856	0,088076416
Recall Avg (15)	0,399484185	0,347772559	0,102677377	0,104421022	0,292186886	0,110701215
Recall Avg (20)	0,484566035	0,463789118	0,143163089	0,1321988	0,352236805	0,133545601

Tabela 4.21 - Recall médio para Português para o Avaliador Prof. Gabriel Lopes

Recall \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0,100914266	0,062085921	0,085279527	0,084681554	0,080534448	0,089853115
Recall Avg (10)	0,166227626	0,135645273	0,155097352	0,158861147	0,137478892	0,146752468
Recall Avg (15)	0,211752786	0,208842305	0,19441078	0,193804458	0,198855961	0,197265355
Recall Avg (20)	0,285856612	0,291097308	0,228846158	0,234336855	0,255690645	0,26465666

Tabela 4.22 - Recall médio para Português para o Avaliador Prof. Joaquim Ferreira da Silva

Pelas mesmas razões já descritas sobre as avaliações por parte destes avaliadores, podemos constatar que as mesmas medidas que tinham melhor precisão total média na avaliação fo Prof. Gabriel Lopes também têm a melhor cobertura. Já no que concerne à cobertura média nos resultados do Prof. Joaquim Ferreira da Silva estes são mais equitativos, pelo que diferenciar claramente é difícil mas a medida *Phi-Square* e *Least Tf-Idf* mostram maior cobertura.

Nas secções 8.8 e 8.14 podemos ver os resultados das coberturas para todas as medidas utilizadas nesta dissertação.

4.3 Língua Inglesa

A análise efectuada para a língua inglesa segue os mesmos moldes que o que foi abordado para a Língua Portuguesa. Será seleccionado um documento que tenha sido avaliador por dois avaliadores de onde serão feitas as leituras dos resultados. Mais importa que referir que os resultados em inglês oferecem a possibilidade de fazer uma comparação com os resultados obtidos para a língua inglesa no trabalho [1].

Apresentamos de seguida as avaliações feitas pelos avaliadores Prof. Joaquim Ferreira da Silva e Prof. Gabriel Lopes. Um documento avaliado por ambos é o EN_32006Q804_01[55]

Para as medidas que foram pedidas para serem avaliadas obrigatoriamente, estes autores obtiveram os valores de precisão que são apresentados para as várias medidas ao longo das próximas secções.

[55] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

4.3.1 Phi-Square

Para o *Phi-Square*, a listagem de termos produzida e que foi apresentado aos avaliadores é a seguinte:

Termos	**Valor da Medida**
governing board	0,016368033116676
governing	0,014533005724990
chairperson	0,010633486245839
bureau	0,006954830301350
director	0,004513219266702
founding regulation	0,004090793192082
founding	0,004090793192082
centre	0,003606283277149
director of the centre	0,003272569769547
voting	0,002891409949613
motion	0,002196500393209
if the chairperson	0,002045295373861
meeting	0,001901388889910
attend	0,001811246676773
members	0,001787372645332
minutes	0,001772238243083
he / she	0,001687973498046
members of the governing	0,001636220104502
members of the governing board	0,001636220104502
unable to attend	0,001636220104502
majority	0,001636220104502
vice-chairpersons	0,001636220104502
meetings of the governing board	0,001636220104502
meetings of the governing	0,001636220104502
development of vocational training	0,001293200838982

Tabela 4.23 - Lista de Termos para a medida Phi-Square para o ficheiro en_32006Q804_01.html

Como podemos constatar pela tabela anterior, esta medida apresenta uma boa distinção de termos pelos seus valores, não obstante ao facto, de neste caso aparecerem 7 termos com o mesmo valor. No capítulo 5 veremos mais alguns exemplos de listagens desta medida para se comprovar a sua eficácia na atribuição de pesos aos termos. Além desta boa distinção, podemos observar pelas tabelas de precisão apresentadas a seguir que os resultados são bons.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,8	0	0,8	0.181818181818182	0.296296296296296
10	0,7	0	0,7	0.318181818181818	0.437500000000000
15	0,6	0,066666667	0,666666667	0.409090909090909	0.486486486486486
20	0,55	0,05	0,6	0.500000000000000	0.523809523809524

Tabela 4.24- Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Phi-Square

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,8	0	0,8	0.133333333333333	0.228571428571429
10	0,6	0,1	0,7	0.200000000000000	0.300000000000000
15	0,466666667	0,2	0,666666667	0.233333333333333	0.311111111111111
20	0,4	0,15	0,55	0.266666666666667	0.320000000000000

Tabela 4.25 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Phi-Square

Como podemos observar nas tabelas anteriores, onde são indicados os valores de precisão para os vários patamares escolhidos, podemos observar que a precisão total dos dois avaliadores anda muito próxima, divergindo somente no patamar de 20, mesmo assim uma divergência de 5 décimas.

O Grau de concordância dos avaliadores nesta medida, para um ficheiro na língua inglesa é de 0.72752 o que dá aproximadamente 72.75% de concordância, isto leva o nível de concordância para o patamar de substancial de acordo com a tabela de concordância apresentada na secção 2.8.3. As matrizes de confusão necessárias para o calculo deste valor são apresentadas na secção 8.15.1

4.3.2 Least Tf-Idf

Termos	Valor da Medida
chairperson	0,029851088353419
governing	0,029590879977958
bureau	0,023731661781725
bureau and the governing	0,023731661781725
governing board and the bureau	0,023731661781725
founding	0,013959801048074
director	0,013267150379297
director and deputy director	0,013267150379297
chairperson or the director	0,013267150379297
centre	0,009292295675709
director of the centre	0,009292295675709
voting	0,008844766919532
members of the governing	0,007828313677225
members	0,007828313677225
chairperson considers that a motion	0,007739171054590
motion may impede the governing	0,007739171054590
motion	0,007739171054590
minutes	0,005706481375529
attend	0,005614391842917
majority of members	0,005583920419229
chairperson and the vice-chairpersons	0,005583920419229
majority	0,005583920419229
vice-chairpersons and members	0,005583920419229
majority of its members	0,005583920419229
vice-chairpersons	0,005583920419229

Tabela 4.26 - Lista de Termos para a medida Least Tf-Idf para o ficheiro en_32006Q804_01.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0	0,6	0.136363636363636	0.222222222222222
10	0,7	0	0,7	0.318181818181818	0.437500000000000
15	0,5333333	0,066666667	0,6	0.363636363636364	0.432432432432432
20	0,5	0,1	0,6	0.454545454545455	0.476190476190476

Tabela 4.27 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Tf-Idf

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,2	0,6	0.068965517241379	0.117647058823529
10	0,4	0,3	0,7	0.137931034482759	0.205128205128205
15	0,333333333	0,266666667	0,6	0.172413793103448	0.227272727272727
20	0,3	0,25	0,55	0.206896551724138	0.244897959183673

Tabela 4.28 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Tf-Idf

Na Tabela 4.26, podemos constatar algumas consequências do operador *Least*. Veja-se o termo mais bem pontuado:

Chairperson	0,029851088353419

Se olharmos um pouco mais a meio da mesma tabela podemos encontrar o seguinte:

director	0,013267150379297
director and deputy director	0,013267150379297
chairperson or the director	0,013267150379297

Que claramente demonstra que o efeito Least, que pode ser visto observando o facto de a multipalavra *"chairperson or the director"*ter assumido o menor valor das suas palavras das extremidades, neste caso "*Chairperson*" e "*director*" (as pontuações na Tabela 4.26).

O Grau de concordância dos avaliadores nesta medida, para um ficheiro na língua inglesa é de 0,4375 o que dá aproximadamente 43.75% de concordância, isto leva o nível de concordância seja classificado como moderado de acordo com a tabela de concordância apresentada na secção 2.8.3. As matrizes de confusão necessárias para o cálculo deste valor são apresentadas na secção 8.15.2

4.3.3 Least Median Rvar

Termos	Valor da Medida
vice-chairpersons	17,000000000000000
simultaneously	14,000000000000000
admissibility	13,000000000000000
countersigned	13,000000000000000
far-reaching	12,000000000000000
appointments	12,000000000000000
ascertained	11,000000000000000
explanation	11,000000000000000
nominations	11,000000000000000
nominations and appointments	11,000000000000000
secretariat	11,000000000000000
scrutineers	11,000000000000000
medium-term	11,000000000000000
vice-chairs	11,000000000000000
precedence	10,000000000000000
indication	10,000000000000000
chairperson	9,488692799006760
chairperson and countersigned	9,488692799006760
substance	9,000000000000000
convening	9,000000000000000
seniority	9,000000000000000
forthwith	9,000000000000000
postponed	9,000000000000000
therefrom	9,000000000000000
deletion therefrom	8,500000000000000

Tabela 4.29 - Lista de Termos para a medida Least Median Rvar para o ficheiro en_32006Q804_01.html

Como podemos constatar, ao observar a Tabela 4.29, as variantes da medida Rvar conseguem apresentar resultados com mais diferenciação entre os termos visto que o valor atribuído pela medida, ao contrário da medida base, tem maior variação. Como se pode constatar na Tabela 8.56 da secção 8.18.1, onde se vê a lista de termos para este mesmo documento para a medida Rvar.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,25	0	0,25	0.045454545454545	0.076923076923077
10	0,125	0,375	0,5	0.045454545454545	0.066666666666667
15	0,1818182	0,363636364	0,545454545	0.090909090909091	0.121212121212121
20	0,1875	0,25	0,4375	0.136363636363636	0.157894736842105

Tabela 4.30 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median Rvar

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,2	0,4	0.033333333333333	0.057142857142857
10	0,4	0,2	0,6	0.133333333333333	0.200000000000000
15	0,466666667	0,266666667	0,733333333	0.233333333333333	0.311111111111111
20	0,4	0,3	0,7	0.266666666666667	0.320000000000000

Tabela 4.31 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Median Rvar

O cálculo da estatística Kappa, nesta medida resultou num valor de 0,296536796536796, o que dá aproximadamente 26.65% de concordância, o que é considerado considerável pela Tabela 2.4. As matrizes de confusão necessárias para o cálculo deste valor são apresentadas na secção 8.15.3.

4.3.4 Least Median MI

Termos	Valor da Medida
vice-chairpersons	63,673145221654230
simultaneously	52,436707829597600
admissibility	48,691228698912056
countersigned	48,691228698912056
far-reaching	44,945749568226520
appointments	44,945749568226520
ascertained	41,200270437540970
explanation	41,200270437540970
nominations	41,200270437540970
nominations and appointments	41,200270437540970
secretariat	41,200270437540970
scrutineers	41,200270437540970
medium-term	41,200270437540970
vice-chairs	41,200270437540970
chairperson	40,800226351661344
chairperson and countersigned	40,800226351661344
precedence	37,454791306855430
indication	37,454791306855430
correspondence	37,056135788244060
substance	33,709312176169890
convening	33,709312176169890
seniority	33,709312176169890
forthwith	33,709312176169890
postponed	33,709312176169890
therefrom	33,709312176169890

Tabela 4.32 - Lista de Termos para a medida Least Median MI para o ficheiro en_32006Q804_01.html

NA mesma medida que o *Least Median Rvar*, também o *Least Median MI* apresenta melhores resultados que a sua medida base. Podemos observar pela tabela anterior uma hierarquização dos resultados, se bem com algumas repetições de pesos, que resulta em parte do operador *Least*. Mas se observarmos a Tabela 8.57, presente na secção 8.18.2, constatamos aí uma atribuição de peso igual a todos os termos.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,25	0	0,25	0.045454545454545	0.076923076923077
10	0,125	0,375	0,5	0.045454545454545	0.066666666666667
15	0,25	0,333333333	0,583333333	0.136363636363636	0.176470588235294
20	0,1875	0,25	0,4375	0.136363636363636	0.157894736842105

Tabela 4.33 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median MI

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,2	0,4	0.034482758620690	0.058823529411765
10	0,4	0,2	0,6	0.137931034482759	0.205128205128205
15	0,533333333	0,2	0,733333333	0.275862068965517	0.363636363636364
20	0,45	0,3	0,75	0.300000000000000	0.360000000000000

Tabela 4.34 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Median MI

O valor Kappa obtido é de 0,258474576271186, o que dá aproximadamente 25.84% de concordância, o que é considerado considerável pela Tabela 2.4. As matrizes de confusão necessárias para o cálculo deste valor são apresentadas na secção 8.15.4.

4.3.5 Least Bubbled Median Phi-Square

Termos	Valor da Medida
chairperson	0,116968348704232
governments	0,075066368633285
governing	0,061417937972688
bureau	0,041728981808101
vice-chairpersons	0,041724438596906
governing board and the bureau	0,034121076651493
founding	0,032726345536657
bureau and the governing	0,030708968986344
vice-chairs	0,026998166150939
motions	0,023633032442703
meetings	0,023119033314776
chairperson considers that a motion	0,020256884950889
motion may impede the governing	0,020256884950889
motion	0,020256884950889
meeting	0,020229154150429
governing the centre between meetings	0,020229154150429
motions that the governing	0,018568811204981
attendance	0,017722382430834
voting	0,017348459697676
chairperson and the vice-chairpersons	0,017180651186961
centre between meetings	0,016113607939238
meetings of the governing	0,015894335403908
vice-chairs of the governing	0,014726272445967
chairperson shall close the meeting	0,014449395821735
attendance at meetings	0,014177905944667

Tabela 4.35 - Lista de Termos para a medida Least Bubbled Median Phi-Square para o ficheiro en_32006Q804_01.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0	0,6	0.136363636363636	0.222222222222222
10	0,6	0	0,6	0.272727272727273	0.375000000000000
15	0,6	0	0,6	0.409090909090909	0.486486486486486
20	0,55	0,05	0,6	0.500000000000000	0.523809523809524

Tabela 4.36 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Phi-Square

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,8	0	0,8	0.010126582278481	0.228571428571429
10	0,6	0,1	0,7	0.015189873417722	0.300000000000000
15	0,533333333	0,133333333	0,666666667	0.266666666666667	0.355555555555556
20	0,5	0,1	0,6	0.333333333333333	0.400000000000000

Tabela 4.37 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Bubbled Median Phi-Square

Dos resultados das avaliações para esta medida, podemos constatar que par a ambos os avaliadores a precisão total obtida para ambos é cima de 0.6, o que se pode considerar como bom.

O valor Kappa obtido é de 0,578651685393258, o que dá aproximadamente 57.86% de concordância, o que é considerado moderado pela Tabela 2.4. As matrizes de confusão necessárias para o cálculo deste valor são apresentadas na secção8.15.5

4.3.6 Least Bubbled Median Rvar

Termos	Valor da Medida
vice-chairpersons	16,999999999999996
simultaneously	14,000000000000000
admissibility	13,000000000000000
countersigned	13,000000000000000
far-reaching	12,000000000000000
ascertained	11,000000000000000
explanation	11,000000000000000
vice-chairs	10,999999999999998
chairperson	9,488692799006760
chairperson and countersigned	9,488692799006760
seniority	9,000000000000000
forthwith	9,000000000000000
postponed	9,000000000000000
precedence	8,655720030369995
deletion	8,000000000000000
absolute majority	8,000000000000000
absolute	8,000000000000000
majority	8,000000000000000
founding	7,999999999999998
chairperson thinks	7,332171708323406
revised	7,000000000000000
besides	7,000000000000000
speaker	7,000000000000000
validly	7,000000000000000
figures	7,000000000000000

Tabela 4.38 - Lista de Termos para a medida Least Bubbled Median Rvar para o ficheiro en_32006Q804_01.html

À semelhança do que acontece com a variante *Least Median Rvar*, também a variante *Least Bubbled Median Rvar* apresenta melhores resultados, em termos da hierarquização de termos pelo peso do que a medida base Rvar.

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,25	0	0,25	0.045454545454545	0.076923076923077
10	0,375	0	0,375	0.136363636363636	0.200000000000000
15	0,3	0	0,3	0.136363636363636	0.187500000000000
20	0,2	0,066666667	0,266666667	0.136363636363636	0.162162162162162

Tabela 4.39 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Rvar

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,2	0,2	0,4	0.033333333333333	0.057142857142857
10	0,3	0,2	0,5	0.100000000000000	0.150000000000000
15	0,266666667	0,4	0,666666667	0.133333333333333	0.177777777777778
20	0,25	0,35	0,6	0.166666666666667	0.200000000000000

Tabela 4.40 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Joaquim Ferreira da Silva para o Least Bubbled Median Rvar

O valor de Kappa obtido é de 0,347826086956521, o que dá aproximadamente 34.83% de concordância, o que é considerado considerável pela Tabela 2.4. As matrizes de confusão necessárias para o cálculo deste valor são apresentadas na secção 8.15.6.

4.4 Leitura de Resultados para a Língua Inglesa

Podemos destacar de imediato que três medidas apresentam um grau de concordância substancial, nomeadamente o *Phi-Square*, o *Least Tf-Idf* e o *Least Bubbled Median Phi-Square*.

Podemos verificar uma semelhança na listagem de termos obtidos pela medida *Least Median Rvar* e *Least Median MI.*

Constata-se também, que estas mesmas medidas *Phi-Square, Least Tf-Idf, Least Median RVar, Least Median MI, Least Bubbled Median Phi-Square e Least Bubbled Median Rvar*, apresentam termos com maior significado semântico que as outras medidas avaliadas, nas quais predominam muito verbos, advérbios, adjectivos ou palavras que não trazem nenhuma pista sobre o conteúdo do assunto do documento, veja-se o caso das variantes da medida Rvar e MI. Nas tabelas de termos avaliados pelos avaliadores para estas medidas, *Least Median Rvar* e *Least Bubbled Median Rvar*, que podem ser encontradas nas secções 8.16 e 8.17, vemos uma predominância de maus descritores.

À semelhança do que aconteceu com a língua Portuguesa, também na língua Inglesa ouve por parte dos avaliadores duas linhas de raciocínio distintas. Por parte do avaliador Prof. Joaquim Ferreira da Silva vemos que a clssificação "*Near Good*

descritptor" é utilizado em mais situações. O que não é observado por parte do avaliador Prof. Gabriel Lopes.

Outra observação que podemos constatar é que as medidas *Rvar* e *MI* fazem aparecer praticamente os mesmos termos, ver tabelas da secção 8.18. Podemos ver que o comportamento destas medidas é idêntico, não diferenciando pelos pesos os termos apresentados.

No que diz respeito à precisão total média obtida para estes avaliadores, podemos observar as seguintes tabelas. Para o avaliador Prof. Gabriel Lopes a amostra de documentos para a média é de nove documentos. Já para o avaliador Prof. Joaquim Ferreira da Silva a amostra é de 5 documentos.

Prec. \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
T. Prec. Avg (5)	0,844444444	0,785185185	0,472222222	0,472222222	0,8	0,524074074
T. Prec. Avg (10)	0,782716049	0,660714286	0,423677249	0,422619048	0,745679012	0,434259259
T. Prec. Avg (15)	0,729466829	0,660541311	0,395983646	0,38015873	0,7000407	0,392572243
T. Prec. Avg (20)	0,686712498	0,677737645	0,347205364	0,338466951	0,653222654	0,403289547

Tabela 4.41 - Precisões Totais médias para Inglês para o Avaliador Prof. Gabriel Lopes

Prec. \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
T. Prec. Avg (5)	0,88	0,8	0,76	0,76	0,88	0,8
T. Prec. Avg (10)	0,88	0,72	0,76	0,78	0,86	0,72
T. Prec. Avg (15)	0,786666667	0,733333333	0,733333333	0,76	0,84	0,746666667
T. Prec. Avg (20)	0,74	0,74	0,71	0,77	0,83	0,74

Tabela 4.42 - Precisões Totais médias para Inglês para o Avaliador Prof. Joaquim Ferreira da Silva

Para ambos, as medidas com a melhor precisão em média são a *Phi-Square* e a *Least Bubbled Median Phi-Square, o* que também se verificou para a língua Portuguesa. Assinala-se que também aí há uma maior concordância entre os avaliadores. A diferenciação de precisões nas outras medidas, deve-se ao facto já mencionado de um avaliador utilizar mais a classificação "Near Good Descriptor".

No que diz respeito à cobertura média, é possível fazer a mesma leitura que foi feita para a precisão total média. Podemos constatar nas tabelas abaixo, que a medida *Phi-Square* e *Least Bubbled Phi-Square* apresentam melhores resultados em média.

Recall \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0,141469168	0,134872012	0,033474497	0,033474497	0,118659513	0,052743101
Recall Avg (10)	0,289430085	0,252416243	0,066413619	0,056751783	0,241410312	0,099584688
Recall Avg (15)	0,356307435	0,362758999	0,096205324	0,08371186	0,340923484	0,118747989
Recall Avg (20)	0,447504494	0,483546939	0,115926344	0,109508125	0,407674418	0,158558013

Tabela 4.43 – Coberturas médias para Inglês para o Avaliador Prof. Gabriel Lopes

Recall \ Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0,11032156	0,085379238	0,074420563	0,068768095	0,095255064	0,094288932
Recall Avg (10)	0,188927434	0,136075757	0,143388847	0,131018627	0,182920437	0,153563385
Recall Avg (15)	0,232007919	0,211789643	0,204761635	0,199411562	0,252938752	0,210683719
Recall Avg (20)	0,291060625	0,271135828	0,240959922	0,24732508	0,311211215	0,271460746

Tabela 4.44 – Coberturas médias para Inglês para o Avaliador Prof. Joaquim Ferreira da Silva

Nos gráficos apresentados nas secções 8.21 e 8.26 podemos ver para cada avaliador, a precisão total para cada documento avaliado pelos avaliadores em relação a precisão total média. Os gráficos apresentados são somente para os limites 5 e 20, e para as duas melhores medidas consideradas pela análise da Tabela 4.41 e da Tabela 4.42. Apresenta-se também os gráficos para a medida *Least Median Rvar*.

4.5 Língua Checa

Dada a especificidade da língua checa, serão somente apresentadas considerações sobre os resultados de um avaliador. Não se calculará por isso valores de estatística Kappa. As listagens de termos que serão apresentadas dizem respeito ao seguinte ficheiro cs_32006D0644.html[56].

4.5.1 Phi-Square

mnohojazyčnost	0,007099977155724
podskupiny	0,005071328357677
mnohojazyčnosti	0,005071328357677
skupiny	0,004070066317448
vysoké úrovni pro mnohojazyčnost	0,004057029128410
skupina	0,003340670032842
oblasti mnohojazyčnosti	0,003042746678425
pozorovatelům	0,002028481007305
odbornou způsobilostí	0,002028481007305
zřízení skupiny na vysoké	0,002028481007305
konzultovat	0,002028481007305
skupinu konzultovat	0,002028481007305
výdaje na zasedání	0,002028481007305
jména členů	0,002028481007305
skupiny nebo podskupiny	0,002028481007305
odborníkům a pozorovatelům	0,002028481007305
skupina na vysoké	0,002028481007305
osm až dvanáct	0,002028481007305
skupině	0,002028481007305
způsobilostí	0,002028481007305
nahrazeni	0,002028481007305
útvary	0,001341325852520
skupiny na vysoké	0,001341325852520
útvary komise	0,001341325852520
odborníkům	0,001275871712364

Tabela 4.45 - Lista de Termos para a medida Phi-Square para o ficheiro cs_32006D0644.html

[56] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006D0644:CS:HTML

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,6	0,4	1	0.500000000000000	0.545454545454545
10	0,5	0,5	1	0.833333333333333	0.625000000000000
15	0,357142857	0,428571429	0,785714286	0.833333333333333	0.500000000000000
20	0,263157895	0,421052632	0,684210526	0.833333333333333	0.400000000000000

Tabela 4.46 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Phi-Square

Umas das coisas que podemos observar pela precisão alcançada pelo *Phi-Square* para o checo é que se mantém com bons resultados, o que vai de encontro ao que aconteceu com esta medida para as outras línguas. Podemos observar que a precisão total é máxima para os limites de 5 e 10. Também a cobertura apresenta bons resultados.

4.5.2 Least Tf-Idf

mnohojazyčnost	0,025845015734672
podskupiny	0,018460725524766
mnohojazyčnosti	0,018460725524766
skupina	0,013619695407680
mnohojazyčnost zřizuje se skupina	0,013619695407680
skupina a její podskupiny	0,013619695407680
skupiny nebo podskupiny	0,012000622357528
skupiny	0,012000622357528
pozorovatelům	0,007384290209906
konzultovat	0,007384290209906
způsobilostí v oblasti mnohojazyčnosti	0,007384290209906
skupině	0,007384290209906
způsobilostí	0,007384290209906
nahrazeni	0,007384290209906
odborníkům	0,006823998624308
odborníkům a pozorovatelům	0,006823998624308
skupina na vysoké	0,006263707038709
vysoké úrovni pro mnohojazyčnost	0,006263707038709
skupiny na vysoké	0,006263707038709
vysoké	0,006263707038709
útvary	0,005966811313056
skupině přidělily příslušné útvary	0,005966811313056
funkčního období nahrazeni	0,005966811313056
funkčního	0,005966811313056
osobně	0,005966811313056

Tabela 4.47 - Lista de Termos para a medida Least Tf-Idf para o ficheiro cs_32006D0644.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,4	0,8	0.333333333333333	0.363636363636364
10	0,2	0,6	0,8	0.333333333333333	0.250000000000000
15	0,2	0,533333333	0,733333333	0.500000000000000	0.285714285714286
20	0,2	0,45	0,65	0.666666666666667	0.307692307692308

Tabela 4.48- Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Tf-Idf

À semelhança da medida anterior, também o *Least Tf-Idf* mostra ter de acordo com a avaliação feita, bons resultados de precisão total, mas perde cobertura em relação ao *Phi-Square*.

4.5.3 Least Median Rvar

mnohojazyčnosti	15,000000000000000
mnohojazyčnost	14,000000000000000
projednávaných	14,000000000000000
pozorovatelům	13,000000000000000
způsobilostí	12,000000000000000
zabezpečuje	11,000000000000000
konzultovat	11,000000000000000
shromažďují	11,000000000000000
zabezpečuje sekretářské	11,000000000000000
sekretářské	11,000000000000000
pozorovatelům cestovní	10,500000000000000
podskupiny	10,000000000000000
prostorách	10,000000000000000
nepřísluší	10,000000000000000
neexistuje	10,000000000000000
zveřejněna	10,000000000000000
podskupiny budou rozpuštěny	10,000000000000000
zveřejňují	10,000000000000000
jednotlivě	10,000000000000000
rozpuštěny	10,000000000000000
způsobilostí v oblasti mnohojazyčnosti	9,500000000000000
důvěrných	9,000000000000000
zveřejnit	9,000000000000000
zůstávají	9,000000000000000
původním jazyce dotyčného dokumentu zveřejnit	9,000000000000000

Tabela 4.49 - Lista de Termos para a medida Least Median Rvar para o ficheiro cs_32006D0644.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,2	0,6	0.333333333333333	0.363636363636364
10	0,2	0,1	0,3	0.333333333333333	0.250000000000000
15	0,133333333	0,133333333	0,266666667	0.333333333333333	0.190476190476190
20	0,1	0,1	0,2	0.333333333333333	0.153846153846154

Tabela 4.50 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median Rvar

4.5.4 Least Median MI

mnohojazyčnosti	72,258388635426410
mnohojazyčnost	67,441162726397980
projednávaných	67,441162726397980
pozorovatelům	62,623936817369554
způsobilostí	57,806710908341130
zabezpečuje	52,989484999312700
konzultovat	52,989484999312700
shromažďují	52,989484999312700
zabezpečuje sekretářské	52,989484999312700
sekretářské	52,989484999312700
pozorovatelům cestovní	50,580872044798490
pozorovatele	49,488944741621780
zpracovávají	49,488944741621780
podskupiny	48,172259090284270
prostorách	48,172259090284270
nepřísluší	48,172259090284270
neexistuje	48,172259090284270
zveřejněna	48,172259090284270
podskupiny budou rozpuštěny	48,172259090284270
zveřejňují	48,172259090284270
jednotlivě	48,172259090284270
rozpuštěny	48,172259090284270
způsobilostí v oblasti mnohojazyčnosti	45,763646135770060
zveřejňování	44,623363444323815
spravováno_úřadem_pro_úřední_tisky	43,889423070017045

Tabela 4.51 - Lista de Termos para a medida Least Median MI para o ficheiro cs_32006D0644.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,2	0,6	0.333333333333333	0.363636363636364
10	0,2	0,1	0,3	0.333333333333333	0.250000000000000
15	0,133333333	0,133333333	0,266666667	0.333333333333333	0.190476190476190
20	0,1	0,1	0,2	0.333333333333333	0.153846153846154

Tabela 4.52 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Median MI

4.5.5 Least Bubbled Median Phi-Square

mnohojazyčnosti	0,168306300320869
mnohojazyčnost	0,157085880299478
podskupiny	0,060856443666432
podskupin	0,054770799299789
skupinou	0,051735406981616
skupina	0,045268481108914
skupinu	0,045268481108914
mnohojazyčnost zřizuje se skupina	0,045268481108914
skupině	0,045268481108914
skupiny	0,045268481108914
skupiny nebo podskupiny	0,042599510566502
skupin	0,038801555236212
skupin a podskupin	0,036513866199859
skupina a její podskupiny	0,033471044016537
zveřejňování	0,011201716547091
skupině přidělily příslušné útvary	0,010730606820159
pozorovatelům	0,010324493133595
nepřísluší	0,010142321146361
neexistuje	0,010142321146361
podskupiny budou rozpuštěny	0,010142321146361
rozpuštěny	0,010142321146361
pozorovatele	0,009530301354087
zveřejnění	0,009334763789243
zveřejněna	0,009334763789243
zveřejňují	0,009334763789243

Tabela 4.53 - Lista de Termos para a medida Least Bubbled Median Phi-Square para o ficheiro cs_32006D0644.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,6	1	0.333333333333333	0.363636363636364
10	0,2	0,7	0,9	0.333333333333333	0.250000000000000
15	0,133333333	0,733333333	0,866666667	0.333333333333333	0.190476190476190
20	0,1	0,6	0,7	0.333333333333333	0.153846153846154

Tabela 4.54 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Phi-Square

À semelhança com a sua medida base, o *Least Bubbled Median Phi-Square* apresenta uma boa precisão total, apesar de perder na cobertura, onde o *Phi-Square* mostra melhores resultados.

4.5.6 Least Bubbled Median Rvar

mnohojazyčnosti	14,570893949858611
mnohojazyčnost	13,599501019868036
podskupiny	10,000000000000000
nepřísluší	10,000000000000000
neexistuje	10,000000000000000
podskupiny budou rozpuštěny	10,000000000000000
rozpuštěny	10,000000000000000
vyzrazeny	9,000000000000000
podskupin	9,000000000000000
podskupiny nesmějí být vyzrazeny	8,000000000000000
nepřísluší odměna	8,000000000000000
nedodrží	8,000000000000000
pozorovatelům	7,709636786377628
zabezpečuje	7,315962630517282
pozorovatele	7,116587802810118
vlivech	7,000000000000000
tématem	7,000000000000000
dodávat	7,000000000000000
dodávat nové podněty a nápady	6,000000000000000
nápady	6,000000000000000
usoudí	6,000000000000000
uhradí	6,000000000000000
limitů	6,000000000000000
zřídit	6,000000000000000
odměna	6,000000000000000

Tabela 4.55 - Lista de Termos para a medida Least Bubbled Median Rvar para o ficheiro cs_32006D0644.html

Threshold	Precision	Precision NearGood	Total Precision	Recall	F-Measure
5	0,4	0,2	0,6	0.333333333333333	0.363636363636364
10	0,2	0,2	0,4	0.333333333333333	0.250000000000000
15	0,133333333	0,266666667	0,4	0.333333333333333	0.190476190476190
20	0,1	0,2	0,3	0.333333333333333	0.153846153846154

Tabela 4.56 - Resultados de Precisão, Cobertura e F-Measure do Avaliador Prof. Gabriel Lopes para o Least Bubbled Median Rvar

Relativamente à medida *Least Median Rvar* (secção 4.5.3) obtiveram-se aqui valores de precisão total ligeiramente mais elevados.

4.6 Leitura de Resultados para a Língua Checa

A amostra de documentos de checo avaliados pelo Prof. Gabriel Lopes é contabilizada em 4 documentos. Nas seguintes tabelas podemos ver a precisão total média e a cobertura média obtida da análise das avaliações deste avaliador.

Prec.\Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
T Prec. Avg (5)	0,7	0,75	0,45	0,45	0,55	0,5
T Prec. Avg (10)	0,7	0,7	0,307142857	0,307142857	0,65	0,386111111
T Prec. Avg (15)	0,625595238	0,642857143	0,229166667	0,245833333	0,633333333	0,31025641
T Prec. Avg (20)	0,605427632	0,607236842	0,221323529	0,221323529	0,5875	0,273529412

Tabela 4.57 - Precisões Totais médias para Checo para o Avaliador Prof. Gabriel Lopes

Recall\Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0,192927171	0,261554622	0,160539216	0,160539216	0,083333333	0,172443978
Recall Avg (10)	0,391981793	0,380077031	0,160539216	0,160539216	0,24947479	0,184348739
Recall Avg (15)	0,421393557	0,474964986	0,172443978	0,172443978	0,389005602	0,184348739
Recall Avg (20)	0,504026611	0,569852941	0,246848739	0,246848739	0,427521008	0,196253501

Tabela 4.58 - Coberturas médias para Checo para o Avaliador Prof. Gabriel Lopes

No seguimento do que aconteceu com as outras línguas utilizadas na experimentação, as medidas com melhores resultados, resultantes, no caso do Checo, da avaliação por parte do avaliador Prof. Gabriel Lopes, são novamente o *Phi-Square*, o *Least Tf-Idf* e o *Least Bubbled Median Tf-Idf*.

Capitulo 5

Conclusões e Trabalho Futuro

5.1 Conclusões

A aposta na extracção de prefixos, que é uma das inovações desta tese, trouxe como consequência a obtenção dos melhores valores de precisão obtidos para as três línguas: 84,6% para o Inglês, 80% para o Checo e 86% para o Português, todas obtidas pela medida *Least Bubbled Tf-Idf* se considerarmos os resultados para o avaliador Prof. Gabriel Lopes (ver Tabela 8.28, Tabela 8.57 e Tabela 8.72). Considerando o avaliador Prof. Joaquim Ferreira da Silva, a precisão para Inglês atingiu 92% para a medida *Least Bubbled Median Tf-Idf* e 84% para Português em quatro medidas, *Least Bubbled Median Tf-Idf, Bubbled Tf-Idf, Tf-Idf* e *Phi-Square* (ver Tabela 8.30 e Tabela 8.61).
Constatámos, maior concordância entre os resultados de dois avaliadores para Português e Inglês nas avaliações feitas nas medidas *Phi-Square, Least Tf-Idf* e *Least Bubbled Median Phi-Square.* No entanto acredito que alguma troca de impressões entre os dois avaliadores relativamente a critérios a utilizar poderia ter aproximado os dois tipos de avaliação. Independentemente disso, parece-me que, com mais tempo teria obtido avaliações de mais pessoas, exigindo a cada um desses avaliadores menos esforço.

Ao filtramos palavras com um comprimento inferior a seis caracteres (este foi um parâmetro utilizado que pode ser alterado, reconfigurando o protótipo construído) e ao termos filtrado multipalavras extraídas que contivessem sinais de pontuação, números e outros símbolos, ao fazer a avaliação dos resultados obtidos sobre a extracção de termos chave, constatámos que a medida *Tf-Idf* não era tão má quanto se dizia em [1].

Bem pelo contrário, os termos chave extraídos com qualquer das variantes desta medida ultrapassam em muito, em valores de precisão, os resultados obtidos utilizando qualquer das variantes da medida *Rvar*, que é considerada a melhor medida em [1].

Mais podemos afirmar, observando as tabelas com os termos extraídos pelo *Rvar* e pela *MI* (secções 8.4, 8.18 e 8.30) que produzem sensivelmente a mesma lista de termos. No que diz respeito à listagem produzida em Checo, as listagens são idênticas para ambas as medidas. Nestas mesmas secções podemos encontrar as listagens para a medida *Tf-Idf*, nas quais podemos constatar que produz resultados visivelmente melhores como já foi dito.

Ambas, *Rvar* e *MI*, sofrem do problema de ser impossível diferenciar pelo peso dos termos qualquer hierarquização de resultados. Além disso parecem escolher termos muito específicos.
As variantes destas medidas, obtidas pelo uso dos operadores "*Least*", "*Bubble*", a conjugação destas duas e o uso da mediana, apresentam melhores resultados, como foi possível ver no caso estudado para as várias línguas ao longo do capítulo 4.

Foi possível verificar ao longo do capítulo 4, na análise dos resultados para as várias medida que a precisão total em média era favorável ao *Phi-Square* e à sua variante *Least Bubbled Median Phi-Square*.

Comparando os valores médios das precisões para o mesmo avaliador (Tabela 4.19, Tabela 4.41 e Tabela 4.57), verifica-se que o Inglês tem a maior precisão assinalada para os primeiros cinco termos extraídos (84,4%) utilizando a medida *Phi-Square* contra 72,8% para o Português e 75% para o Checo mas utilizando a medida *Least Tf-Idf*.

A utilização das Suffix Arrays mostrou-se bastante produtiva nos tempos de extracção das palavras e prefixos desta estrutura. Questões de performance neste caso foram totalmente alcançadas. Existe um problema a ser optimizado de futuro que é a incorporação do extractor de multipalavras como parte integrante do sistema.

5.2 Trabalho Futuro

Sendo o principal objectivo do trabalho apresentado a ordenação de palavras-chave, através de medidas para a extracção de palavras e/ou multipalavras que sejam considerados como bons descritores de documentos, antevemos uma possível futura utilização deste trabalho nas áreas de agrupamento e classificação de documentos.

O trabalho realizado nesta tese possibilitou a criação de várias medidas (ver secção 3.2) que poderiam ser utilizadas numa adaptação do trabalho realizado por de David Ferreira [12]. Adaptação que consistiria em experimentar uma das medidas criadas neste trabalho no seu cálculo da importância de um termo.

Em alemão, onde os nomes podem resultar da concatenação de vários elementos, correspondendo também a nomes compostos ou multipalavras. A extracção de sequências de 4 ou 5 caracteres (não necessariamente prefixos) que faríamos borbulhar (*Bubbling*) de forma análoga à utilizada com os prefixos, poderá ser altamente produtiva. Se, pretendêssemos estender a metodologia desenvolvida nesta dissertação, bem como a aplicação de todas as medidas desenvolvidas a línguas orientais, como o Chinês ou o Japonês, trabalharíamos provavelmente com sequências de 2 caracteres, eventualmente 3, ou mesmo um único carácter porque, nestas línguas, não existe o espaço em branco como separador de palavras e porque há palavras de conteúdo que se escrevem com um único carácter. Aí, a extracção de multi-caracteres correspondestes a conceitos pode ser feita utilizando a mesma maquinaria que utilizei para a extracção de multipalavras. A técnica de *"Bubbling"* é que não seria aplicável.

É Possível fazer a adaptação do protótipo resultante do trabalho realizado na Tese para uma ferramenta de produção com enormes potencialidades a nível científico, para análise de resultados deste tipo de experimentação.
É possível que um trabalho futuro seja o de estudar o uso de outras estruturas de dados além das Suffix Arrays para usar na extracção de termos de documentos.

Estão em progresso trabalhos de escrita de artigos científico baseados nos resultados obtidos nesta dissertação para poderem passar nos testes de *Peer Review*.

Anexo 1 – Módulos de código

6.1 Fiheiros JNI

6.1.1 Header File

Ficheiro header criado pelo comando javah.

```
/* DO NOT EDIT THIS FILE - it is machine generated */
#include <jni.h>
/* Header for class sufArray_SuffixArray */

#ifndef _Included_sufArray_SuffixArray
#define _Included_sufArray_SuffixArray
#ifdef __cplusplus
extern "C" {
#endif
/*
 * Class:     sufArray_SuffixArray
 * Method:    jsarrayString
 * Signature: (Ljava/lang/String;[II)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jsarrayString
  (JNIEnv *, jclass, jstring, jintArray, jint);

/*
 * Class:     sufArray_SuffixArray
 * Method:    jlcp
 * Signature: ([ILjava/lang/String;[II)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jlcp
  (JNIEnv *, jclass, jintArray, jstring, jintArray, jint);

/*
 * Class:     sufArray_SuffixArray
 * Method:    jsuffixsort
 * Signature: ([I[IIII)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jsuffixsort
  (JNIEnv *, jclass, jintArray, jintArray, jint, jint, jint);

#ifdef __cplusplus
}
#endif
#endif
```

6.1.2 Code File

Ficheiro C que implementa o header apresentado na secção anterior.

```
/* DO NOT EDIT THIS FILE - it is machine generated */
#include <jni.h>
/* Header for class SuffixArray */
#ifndef _Included_SuffixArray
#define _Included_SuffixArray
#ifdef __cplusplus
extern "C" {
#endif
/*
 * Class:     sufArray_SuffixArray
 * Method:    jsarrayString
 * Signature: (Ljava/lang/String;[II)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jsarrayString
      (JNIEnv *env, jclass junk, jstring s0, jintArray a0, jint n){

      const jbyte *s = (*env)->GetStringUTFChars(env, s0, 0);
      jint *a = (*env)->GetIntArrayElements(env, a0, 0);
      int r = bsarray(s, a, n);
      (*env)->ReleaseStringUTFChars(env, s0, s);
      (*env)->ReleaseIntArrayElements(env, a0, a, 0);
}
/*
 * Class:     sufArray_SuffixArray
 * Method:    jlcp
 * Signature: ([ILjava/lang/String;[II)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jlcp
  (JNIEnv *env, jclass junk, jintArray a0, jstring s0, jintArray b0, jint
n){
      const jbyte *s = (*env)->GetStringUTFChars(env, s0, 0);
      jint *a = (*env)->GetIntArrayElements(env, a0, 0);
      jint *b = (*env)->GetIntArrayElements(env, b0, 0);
      lcpa(a, s, b, n);
      (*env)->ReleaseStringUTFChars(env, s0, s);
      (*env)->ReleaseIntArrayElements(env, a0, a, 0);
      (*env)->ReleaseIntArrayElements(env, b0, b, 0);
}
/*
 * Class:     sufArray_SuffixArray
 * Method:    jsuffixsort
 * Signature: ([I[IIII)V
 */
JNIEXPORT void JNICALL Java_sufArray_SuffixArray_jsuffixsort
   (JNIEnv * env, jclass junk, jintArray a1, jintArray b1, jint n, jint k,
jint l){

      jint *a = (*env)->GetIntArrayElements(env, a1, 0);
      jint *b = (*env)->GetIntArrayElements(env, b1, 0);
      suffixsort(a,b,n,k,l);
      (*env)->ReleaseIntArrayElements(env, a1, a, 0);
      (*env)->ReleaseIntArrayElements(env, b1, b, 0);
}#ifdef __cplusplus
}
#endif
#endif
```

6.2 Construção da Estrutura de palavras

```
    /**
     *
     * @param saIn
     * @return
     */
    public static Terms
buildSuffixArray_HashMapOf_FullWords_docIndex(SuffixArray saIn, int
wordsLen){
    Terms toReturn = new Terms();
    System.out.println("buildSuffixArray_HashMapOf_FullWords_docIndex");
    //Cycle to set the partial of the terms by document.
    for(int i = 0 ; i < saIn.a.length; i++)
    {
      String suffix = saIn.s.substring(saIn.a[i]);
     int suffixPos = saIn.a[i];
      if(suffix.startsWith(" "))
      {
       if (suffix.length() > 1)
       {
        //Suffixes starting by numbers or by symbols are not considered
        if ((saIn.isDigit(suffix) == false) && (saIn.isSymbol(suffix) ==
false))
        {
         int toIndex = suffix.indexOf(" ", 1);
         if( toIndex > 0)
         {
          //Palavras com comprimento maior do que 6 Ã© que serao inseridas
na HashMap de termos.
          //( 6 = 5 caracteres + espaÃ§o no inicio)
          if (toIndex > wordsLen)
          {
           String word = suffix.substring(0, toIndex);
           for (Document doc : documents) {
            if (doc.belongsToDocument(suffixPos)) {
             toReturn.insertNewTerm(suffix.substring(0, toIndex),doc);
             //Criar um HashMap onde vou guardar Termo , contador para o
numero de vezes que o termo occore.
             //Guardar tambem em que documentos ocorre.
            }
           }
          }
         }
        }
       }
      }
     return toReturn;
    }
```

6.3 Construção da Estrutura de Prefixos

```
     /**
     *
     * @param saIn
     * @param numberofChars
     */
     public static Terms
buildSuffixArray_HashMapOf_Prefixes_docIndex(SuffixArray saIn, int
numberofChars)
     {
     System.out.println("buildSuffixArray_HashMapOf_Prefixes_docIndex");
     Terms Preffix_chars = new Terms();
     int innernumberofChars = numberofChars + 1;
     for(int i = 0 ; i < saIn.a.length; i++)
     {
       String suffix = saIn.s.substring(saIn.a[i]);
       int suffixPos = saIn.a[i];
       if(suffix.startsWith(" "))
       {
          if (suffix.length() > 1)
          {
            //Prefixes starting by numbers or by symbols are not considered
            if ((saIn.isDigit(suffix) == false) && (saIn.isSymbol(suffix)
== false))
            {
               int toIndex =  innernumberofChars;
               String preffix = suffix.substring(0, toIndex);
               if (!(preffix.trim().length() < numberofChars ))
               {
                if (!preffix.trim().contains(" ")){
                for (Document doc : documents)
                {
                   if (doc.belongsToDocument(suffixPos))
                   {
                     Preffix_chars.insertNewTerm(preffix, doc);
                   }
                 }
              }
          }
       }
      }
     }
    }
     return Preffix_chars;
    }
```

Anexo 2 – Manual do Utilizador do Protótipo.

O Protótipo desenvolvido é composto por três componentes diferentes. Uma primeira janela apresenta todas as configurações possíveis que se podem aplicar para obter os resultados.

As outras duas componentes são a janela de avaliação de termos e a outra a janela de leitura das avaliações feitas pelos avaliadores.

7.1 Janela de Configuração

A seguinte Figura 7.1 disponibiliza ao utilizador todos os parâmetros de configuração possíveis de alterar. Ao longo desta secção detalhar-se-á os diversos componentes, nomeando a sua funcionalidade.

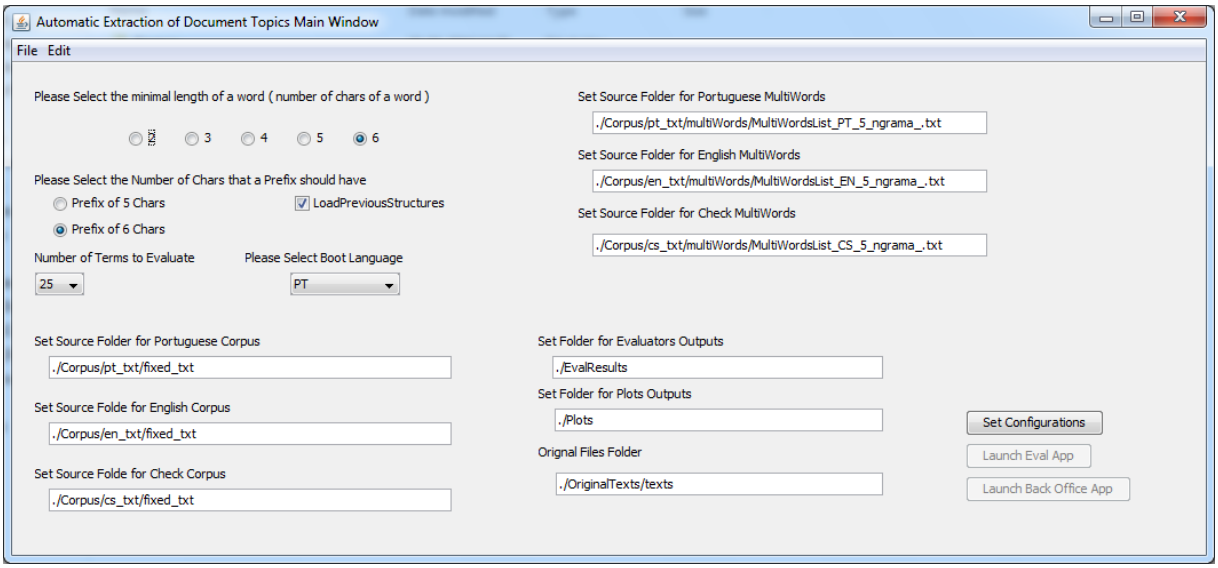


Figura 7.1 - Janela de Configuração

A seguinte figura, permite ao utilizador definir qual o tamanho mínimo de caracteres que uma palavra deve ter.

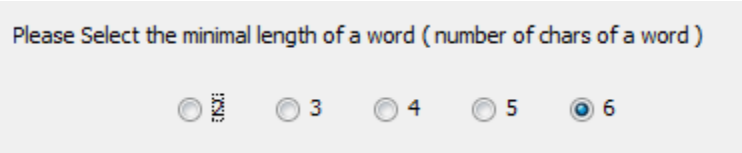


Figura 7.2 - Componente de selecção do comprimento de caracteres mínimo de uma palavra

Já na Figura 7.3, podemos ver a opção de escolher o tamanho que os prefixos devem ter e uma opção para fazer o carregamento das estruturas de dados previamente utilizadas. Se porventura, o utilizador desejar usar alguma configuração que seja diferente da última que utilizou, esta opção não deverá ser utilizada.

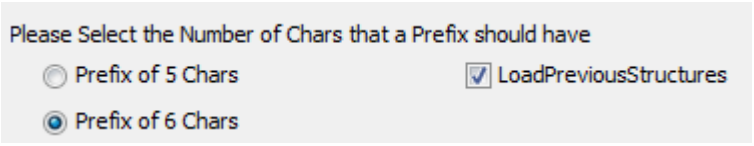


Figura 7.3 - Selecção do tamanho dos Prefixos, e se a aplicação deve carregar as estruturas anteriores ou não.

Nas Figuras 7.4 e 7.5 é possível ver como se selecciona o número de termos que o avaliador terá para avaliar. No trabalho desenvolvido nesta tese, o número de termos utilizado foi de 25.

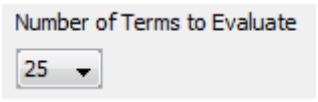


Figura 7.4 - Componente de selecção do numero de termos para avaliar

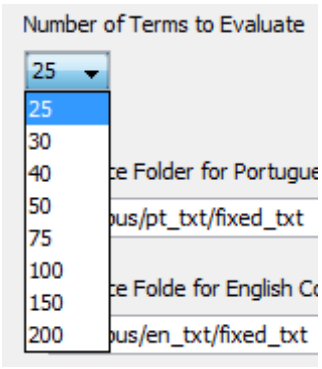


Figura 7.5 - Componente de selecção do numero de termos para avaliar expandido.

Já nas seguintes figuras, podemos ver como se selecciona a língua de arranque das outras duas componentes do protótipo, ver secções 7.2 e 7.3.

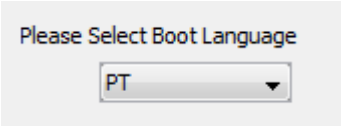


Figura 7.6 - Compoente de selecção da lingua de arranque das aplicações

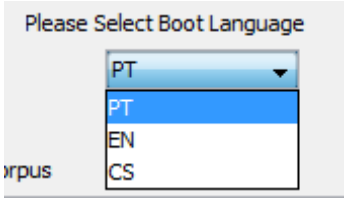


Figura 7.7 - Componente de selecção da língua de arranque das aplicações expendida**.**

As Figuras 7.8, 7.9 e 7.10, servem para o utilizador configurar as pastas onde estão localizados os textos necessários para o funcionamento do protótipo.

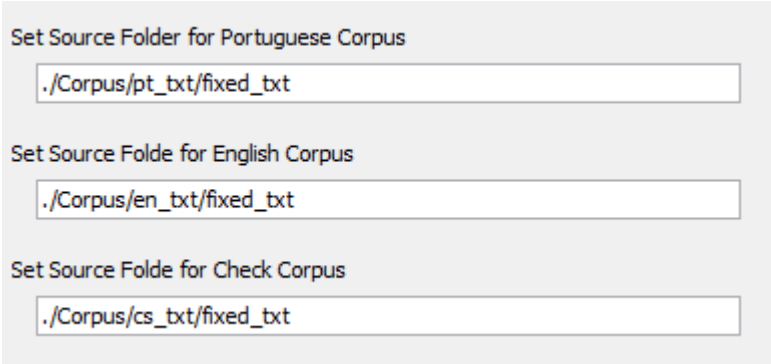


Figura 7.8 - Componentes onde se define a localização dos textos que farão parte do corpus nas diferentes línguas.

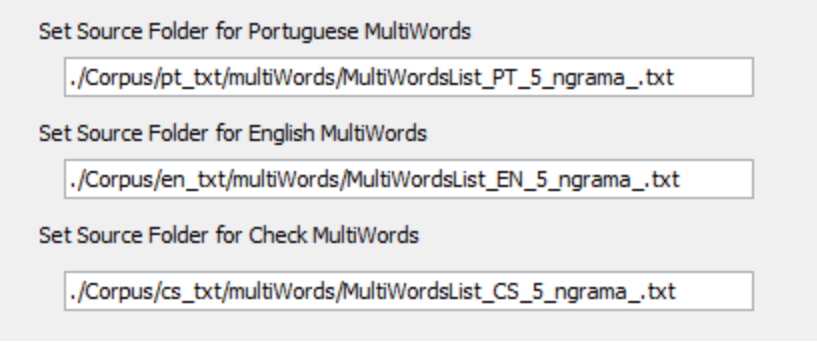


Figura 7.9 - Componentes onde se define a localização dos ficheiros com as multipalavras dos textos tratados das diferentes línguas.

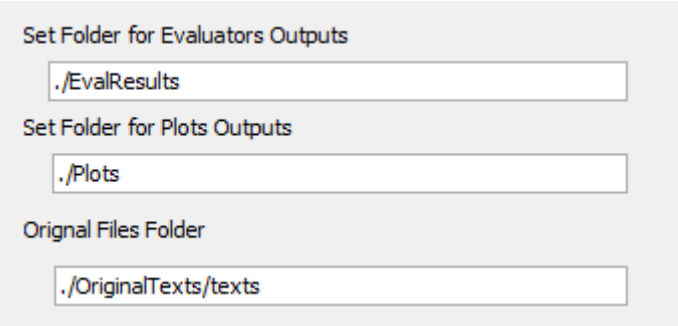


Figura 7.10- Componentes de configuração das pastas de output, e localização dos textos originais

Tendo o utilizador configurado o que ache necessário, terá de fazer *"Set Configurations"*. Após isso, haverá uma transição de estado dos botões que lançam as outras duas componentes do protótipo.

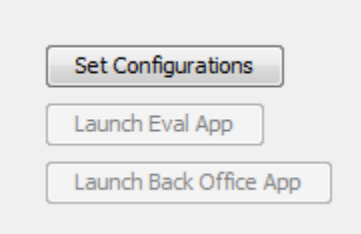


Figura 7.11 - Botão que faz o "Set" das configurações prentendidas, desbloqueando ou outros botões ver Figura 7.12

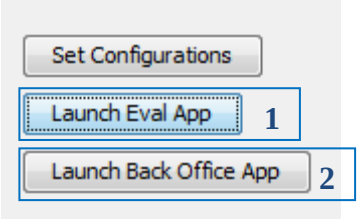


Figura 7.12 - Botões que lançam a Aplicação para os Avaliadores o a Aplicação de "BackOffice"

Na Figura 7.12, o botão identificado por (1) lançará a aplicação de avaliação de termos, ver secção 7.2. Já o botão identificado por (2) lançará o *backOffice*, aplicação que serve para fazer uma análise sobre os resultados das avaliações dos termos por parte dos avaliadores, ver secção 7.3 para mais informação.

7.2 Janela de Avaliação de Termos

Segue-se de seguida a explicação detalhada da janela apresentada aos avaliadores para estes poderem avaliar os termos de cada documento.

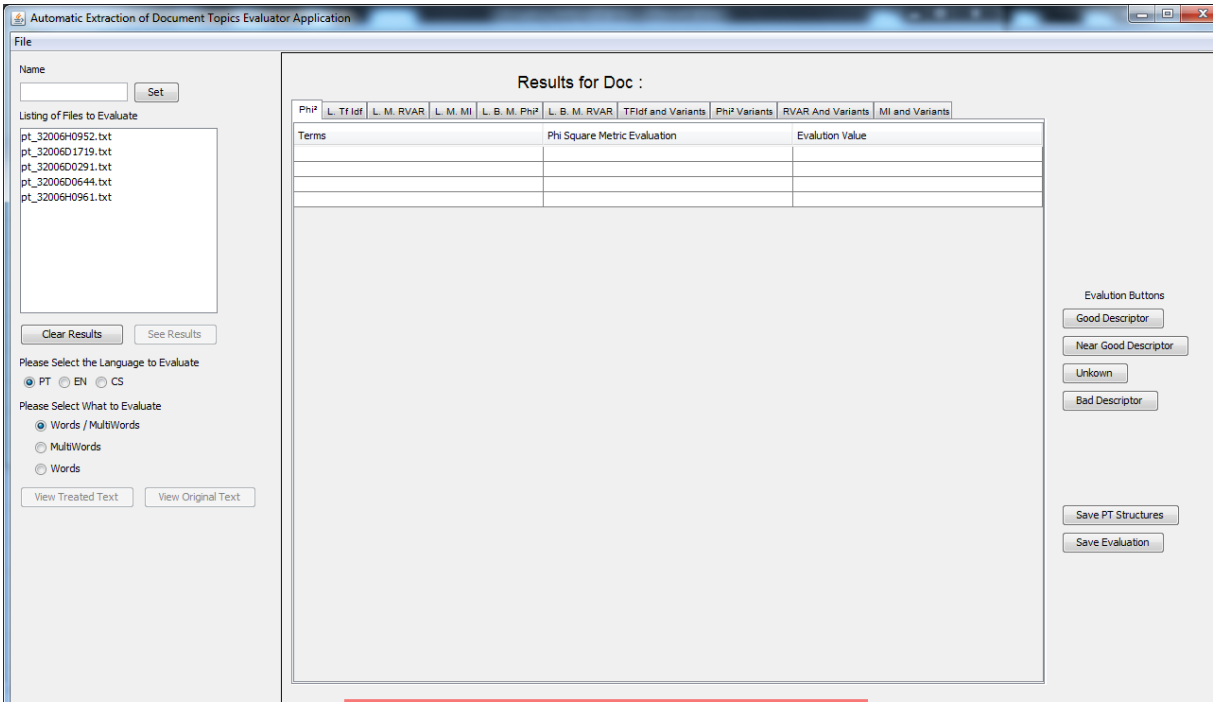


Figura 7.13 – Janela da aplicação dos avaliadores.

A primeira coisa que é pedida a um avaliador é que se identifique. Um exemplo pode ser visto na sequência de Figuras, 7.14 e 7.15.

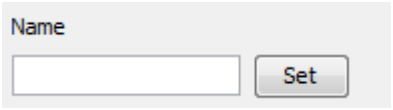


Figura 7.14 – Componente para o avaliador se identificar

Figura 7.15 - Componente onde o avaliador se identificou

Ao fazer p “Set” do seu nome o avaliador desbloqueará o botão, ver Figuras 7.16 e 7.17, que o irá permitir ver os resultados para um determinado documento seleccionado, ver Figura 7.18.

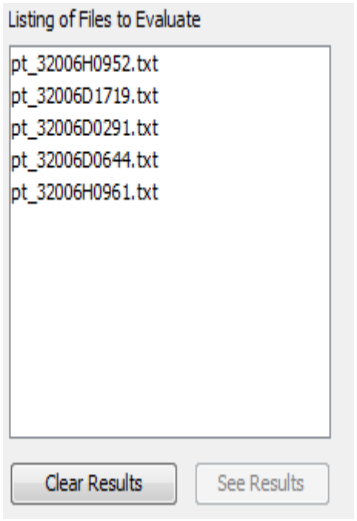


Figura 7.16 – Componente com Lista Inicial de documentos

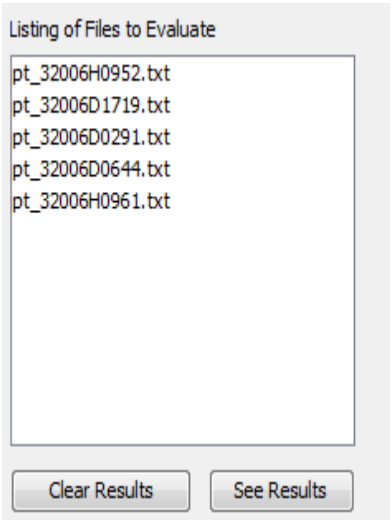


Figura 7.17 - Componente com Lista Inicial de documentos, botão "See Results" activo

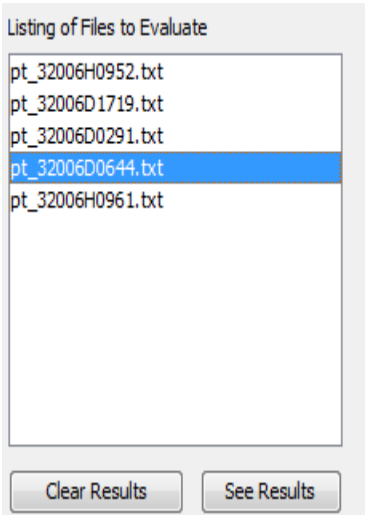


Figura 7.18 - - Componente com Lista Inicial de documentos, com um documento seleccionado

Estando um avaliador no estado presente na Figura 7.18, ao clicar no botão "See Results", o avaliador verá listagem de termos para o documento seleccionado, como podemos ver na Figura 7.24. Ao clicar no botão "Clear Results" o avaliador irá limpar a tabela de resultados voltando ao estado inicial, como se pode ver na Figura 7.23.
Ao clicar num documento, o avaliador vai desbloquear os botões que permitem ver o conteúdo dos documentos, ver Figura 7.22, que inicialmente estão bloqueados como se pode ver na Figura 7.21.

Na Figura 7.19 é onde é possível a um avaliador mudar a língua dos documentos que está avaliar. Se mudar para EN, a listagem apresentada na Figura 7.16 será populada com os documentos em inglês que foram processados pelo protótipo.

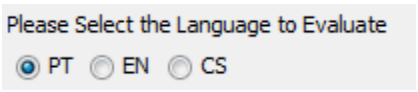


Figura 7.19 – Componente para mudar a língua dos documentos a avaliar.

Já na Figura 7.20, é oferecida a possibilidade de o avaliador ver os resultados só com palavras, ou só com multipalavras. Mas a avaliação de resultados só é permitida para palavras e multipalavras em simultâneo. Essa avaliação é feita utilizando os botões apresentados na Figura 7.25.

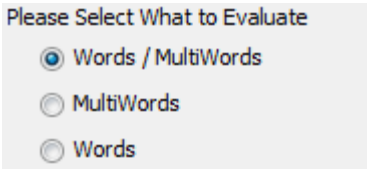


Figura 7.20 – Componente para escolher que tipo de resultados ver (Palavras, Multipalavras ou Ambos)

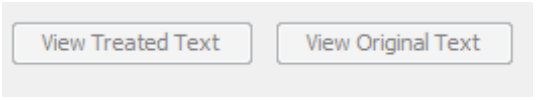


Figura 7.21 – Botões para ver o texto do documento, tratado ou original

Figura 7.22 - Botões para ver o texto do documento, tratado ou original, activos.

Na figura seguinte podemos ver a tabela onde os termos serão apresentados para serem avaliados, como podemos ver na Figura 7.24.

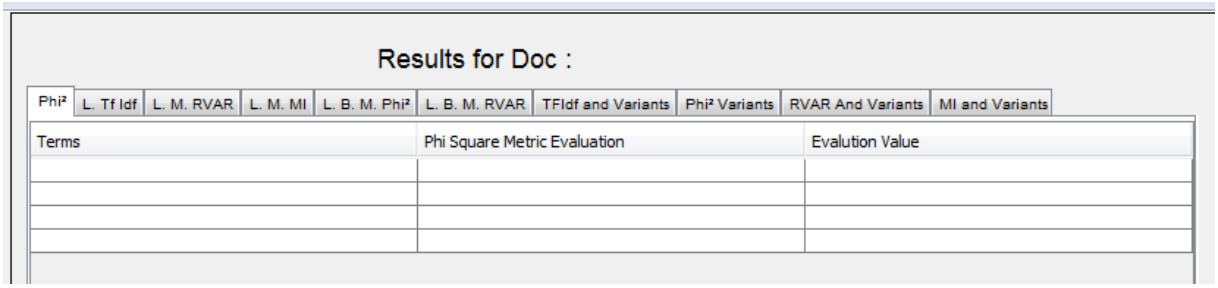
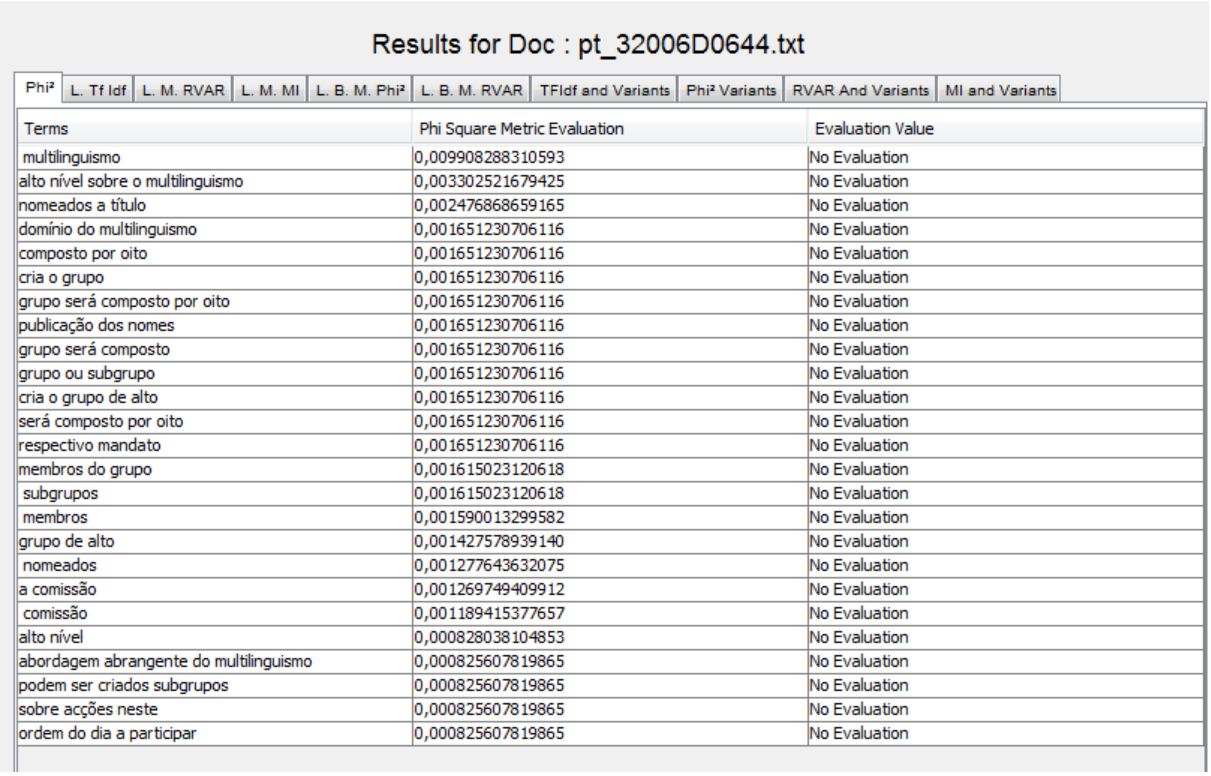


Figura 7.23 – Componente com "tabs", onde vão aparecer as listagens de termos, para as várias medidas.



Terms	Phi Square Metric Evaluation	Evaluation Value
multilinguismo	0,009908288310593	No Evaluation
alto nível sobre o multilinguismo	0,003302521679425	No Evaluation
nomeados a título	0,002476868659165	No Evaluation
domínio do multilinguismo	0,001651230706116	No Evaluation
composto por oito	0,001651230706116	No Evaluation
cria o grupo	0,001651230706116	No Evaluation
grupo será composto por oito	0,001651230706116	No Evaluation
publicação dos nomes	0,001651230706116	No Evaluation
grupo será composto	0,001651230706116	No Evaluation
grupo ou subgrupo	0,001651230706116	No Evaluation
cria o grupo de alto	0,001651230706116	No Evaluation
será composto por oito	0,001651230706116	No Evaluation
respectivo mandato	0,001651230706116	No Evaluation
membros do grupo	0,001615023120618	No Evaluation
subgrupos	0,001615023120618	No Evaluation
membros	0,001590013299582	No Evaluation
grupo de alto	0,001427578939140	No Evaluation
nomeados	0,001277643632075	No Evaluation
a comissão	0,001269749409912	No Evaluation
comissão	0,001189415377657	No Evaluation
alto nível	0,000828038104853	No Evaluation
abordagem abrangente do multilinguismo	0,000825607819865	No Evaluation
podem ser criados subgrupos	0,000825607819865	No Evaluation
sobre acções neste	0,000825607819865	No Evaluation
ordem do dia a participar	0,000825607819865	No Evaluation

Figura 7.24 - – Componente com "tabs", onde vão aparecer as listagens de termos, para as várias medidas, populada.

Tendo os termos disponíveis para serem avaliados, e tendo em conta o conteúdo dos documentos o que é pedido ao avaliador é que classifique os diversos termos apresentados, na escala fornecida pelos botões identificados na Figura 7.25.

A Escala é composta por 4 níveis:

- Good Descriptor
  - Se demonstra o conteúdo do Documento.

- Near Good Descriptor
  - Se dá uma pista sobre o conteúdo do Documento, mas falta algo mais para dar uma ideia mais concreta.

- Bad Descriptor
  - Se for adjectivo marca-se como Bad Descriptor.
  - Se contiver uma forma verbal também deverá ser marcado como Bad Descriptor.
  - Se for um advérbio também deverá sem marcado como Bad Descriptor.

- Unkown
  - Se tiver nomes próprios mencionados no texto, deverá marcar como Unkown.
  - Ou se não souber se de facto descreve o conteúdo.

Esta avaliação pede-se que seja feita na totalidade, para 6 medidas, nomeadamente as identificadas na Figura 7.27, nomeadamente:

- Phi^2 ;
- L.Tf Idf;
- L.M. RVAR;
- L.M. MI;
- L.B.M. Phi^2;
- L.B.M. RVAR:

Figura 7.25 – Botões de Avaliação de Termos

Results for Doc : pt_32006D0644.txt

Phi² | L. Tf Idf | L. M. RVAR | L. M. MI | L. B. M. Phi² | L. B. M. RVAR | TFIdf and Variants | Phi² Variants | RVAR And Variants | MI and Variants

Terms	Phi Square Metric Evaluation	Evaluation Value
multilinguismo	0,009908288310593	Good Topic Descriptor
alto nível sobre o multilinguismo	0,003302521679425	Near Good Descriptor
nomeados a título	0,002476868659165	No Evaluation
domínio do multilinguismo	0,001651230706116	Good Topic Descriptor
composto por oito	0,001651230706116	Bad Descriptor
cria o grupo	0,001651230706116	Bad Descriptor
grupo será composto por oito	0,001651230706116	No Evaluation
publicação dos nomes	0,001651230706116	No Evaluation
grupo será composto	0,001651230706116	No Evaluation
grupo ou subgrupo	0,001651230706116	Unkonwn
cria o grupo de alto	0,001651230706116	No Evaluation
será composto por oito	0,001651230706116	No Evaluation
respectivo mandato	0,001651230706116	No Evaluation
membros do grupo	0,001615023120618	No Evaluation
subgrupos	0,001615023120618	No Evaluation
membros	0,001590013299582	No Evaluation
grupo de alto	0,001427578939140	No Evaluation
nomeados	0,001277643632075	No Evaluation
a comissão	0,001269749409912	No Evaluation
comissão	0,001189415377657	No Evaluation
alto nível	0,000828038104853	No Evaluation
abordagem abrangente do multilinguismo	0,000825607819865	No Evaluation
podem ser criados subgrupos	0,000825607819865	No Evaluation
sobre acções neste	0,000825607819865	No Evaluation
ordem do dia a participar	0,000825607819865	No Evaluation

Figura 7.26 – Tabela de termos com alguns já avaliados.

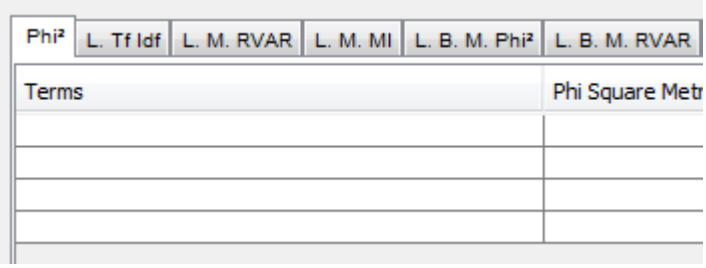


Figura 7.27 – Lista de medidas que são obrigatórias de avaliar.

Os botões da figura seguinte permitem guardar em disco as avaliações feitas pelo avaliador para determinado documento, botão "Save Evaluation". Enquanto que o botão "Save PT Structures" serve para guardar em disco as estruturas utilizadas pelo protótipo, neste caso estruturas de Português. Consoante a língua que estiver a ser avaliada, o botão identificará a língua pela sua abreviatura. "Save EN Structures" para o inglês e "Save CZ Structures" para o Checo.

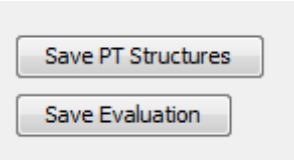


Figura 7.28 – Botões para salvar a Avaliação Efectuada, e o botão para salvar as estruturas de termos criadas.

7.3 Janela de Leitura das Avaliações feitas pelos Avaliadores

Esta, Figura 7.29, é a componente do protótipo que vai permitir fazer a leitura das avaliações feitas pelos avaliadores para os vários documentos. Ao longo desta secção descreve-se os vários componentes oferecidos ao utilizador.

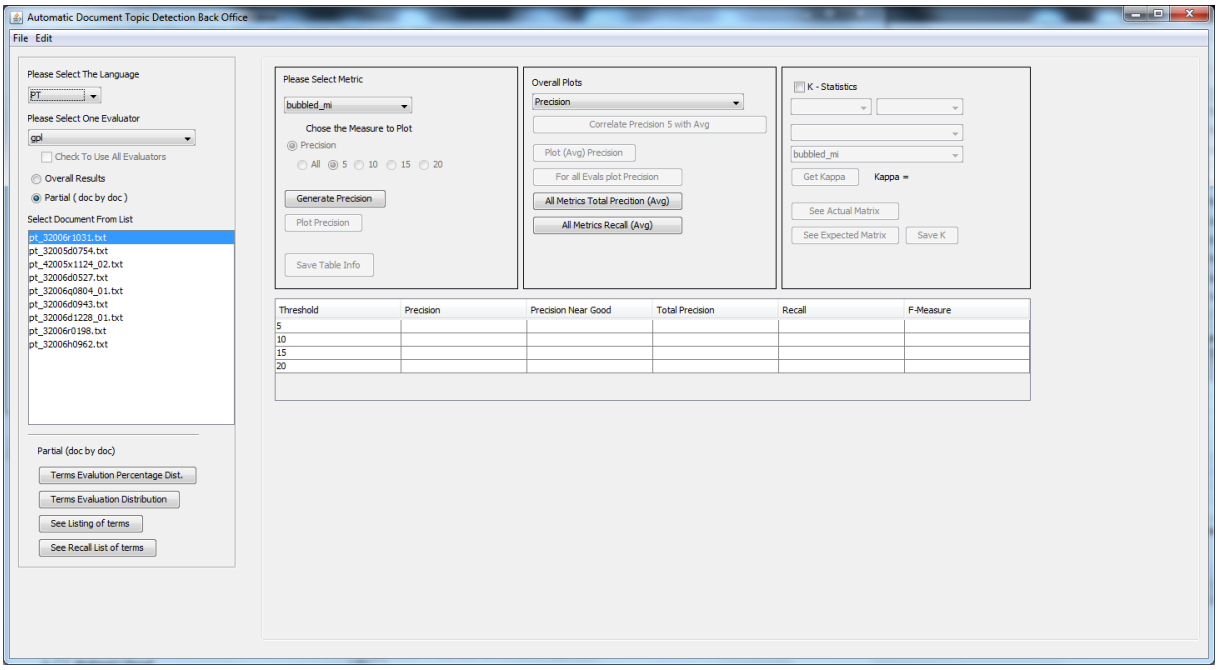


Figura 7.29- Janela da Aplicação de "BackOffice.

À semelhança da componente anterior, também é disponibilizado um componente para alterar a língua sobre a qual se quer ver os resultados das avaliações.

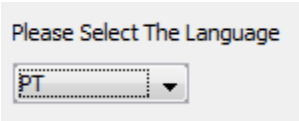


Figura 7.30 – Componente para selecção da língua dos documentos.

Nesta componente do protótipo, podemos escolher, no componente identificado por (1) na Figura 7.31, qual o avaliador de quem queremos ver os resultados. Já a componente identificada por (2) permite alterar a forma como vemos os resultados da avaliação do avaliador. Se parcialmente, documento a documento, se de uma forma total, permitindo a análise da média dos resultados.

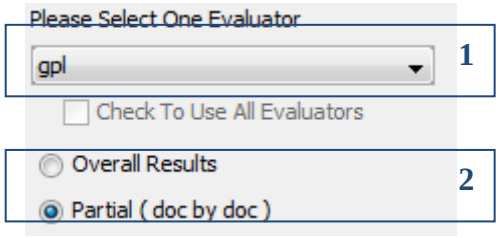


Figura 7.31 – Componente para escolher o avaliador, e componente se avaliação parcial ou total.

A escolha dos documentos é feita recorrendo à lista apresentada na Figura 7.32.

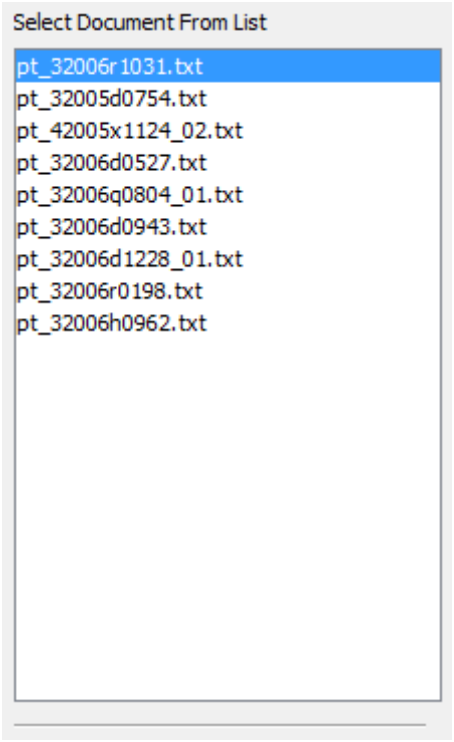


Figura 7.32 Listagem de documentos avaliados pelo avaliador.

Os botões apresentados na seguinte Figura 7.33 permitem obter gráficos e listagens. O botão "Terms Evaluation Percentage Dist" permite visualizar um gráfico como o que pode ser observado na Figura 7.34, que para um determinado documento e uma determinada medida, mostra a percentagem de termos por tipo de avaliação que foi feita pelo avaliador. Já o botão "Terms Evaluation Distribution" apresenta um gráfico como o apresentado na Figura 7.35. Os outros dois botões permitem ver listagens: uma dos termos avaliados pelo autor para determinado documento e para determinada medida. A outra listagem é a que serve de cálculo do Recall para o documento e medida em causa.

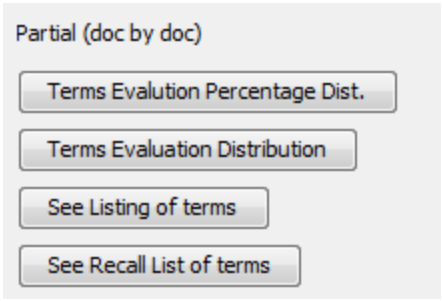


Figura 7.33- Botões que permitem ver a distribuição das avaliações dos autores, e listagens dos termos avaliados.

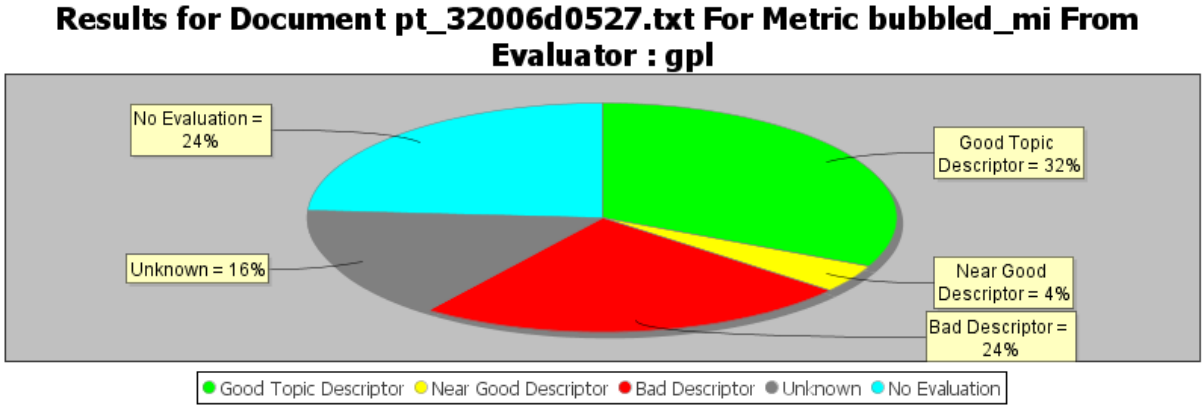


Figura 7.34 – Gráfico exemplificativo

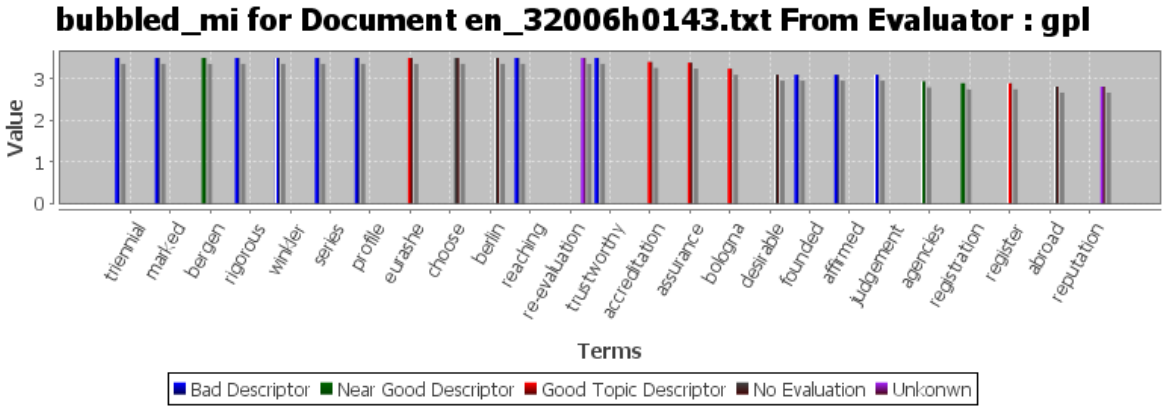


Figura 7.35 - Gráfico exemplificativo

Na próxima sequencia de figuras, 7.36 a 7.38, podemos ver como se selecciona uma medida da qual se queiram ver os resultados. Ao clicar no botão “Generate Precision” a tabela apresentada na Figura 7.44 passa a conter os resultados da precisão (Precision), da cobertura (recall) e da F-Measure, como podemos ver na Figura 7.45.

Ao fazer a geração das medidas, é desbloqueado o botão “Plot Precision” que permite fazer o gráfico da precisão, como se vê na Figura 7.38. Um gráfico exemplificativo é apresentado na Figura 7.39.

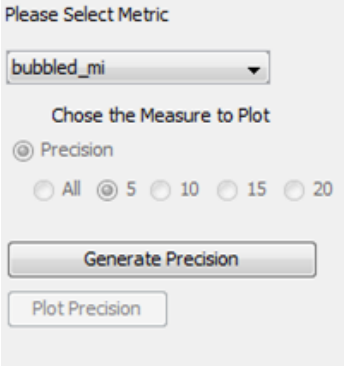


Figura 7.36 – Componente de Selecção da medida.

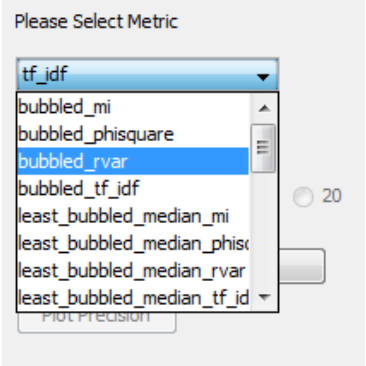


Figura 7.37 - Componente de Selecção da medida expandida

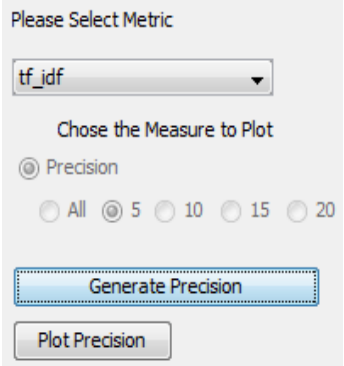


Figura 7.38 – Botões para gerar a Precisão e fazer o gráfico da precisão.

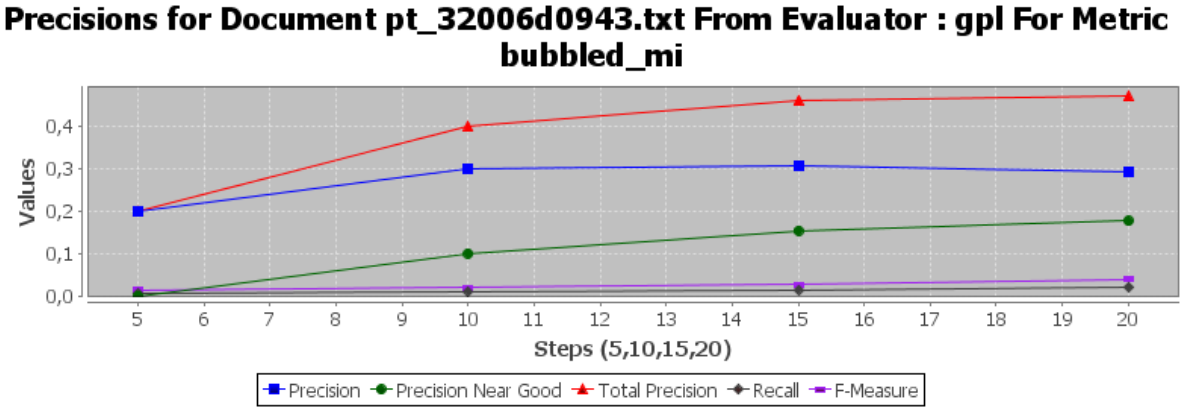


Figura 7.39 – Gráfico exemplo de precisões para um documento e uma determinada medida.

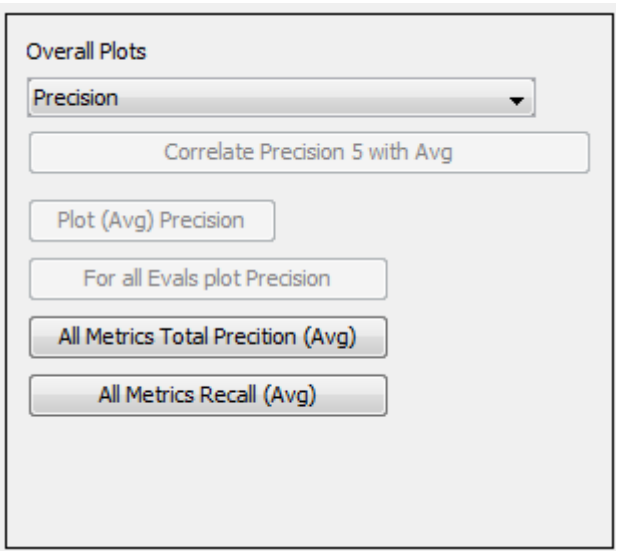


Figura 7.40 – Componente que permite fazer gráficos a correlacionar precisões com a média das precisões.

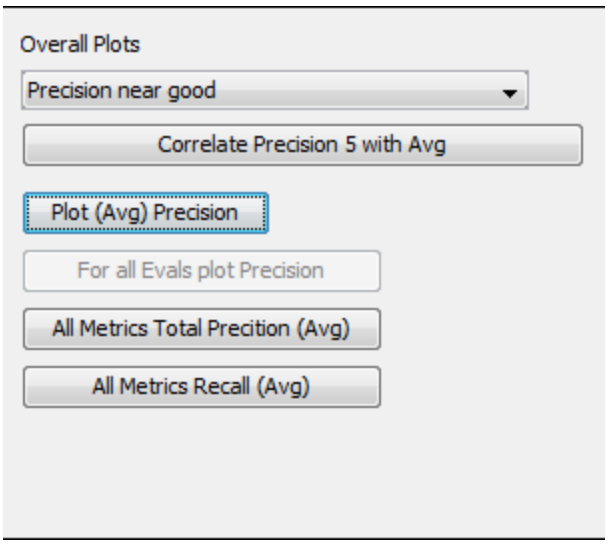


Figura 7.41 - – Componente que permite fazer gráficos a correlacionar precisões com a média das precisões

As duas figuras anteriores permitem fazer gráficos que relacionam a precisão de um determinado documento com a média da precisão, de um determinado avaliador, como podemos ver na Figura 7.43. Permite ainda fazer o gráfico que mostra, para um mesmo documento e medida, qual o valor de precisão e cobertura para um dado limite (5,10,15 ou 20). Um gráfico exemplificativo pode ser vista na Figura 7.42. Já os botões "All Metrics Total Precition (Avg)" e "All Metrics Recall (Avg)", permitem visualizar uma tabela com as precisões totais médias, ou com a cobertura média, para um determinado documento e para todas as medidas avaliadas, para um avaliador. Ver Figura 7.46 e Figura 7.47

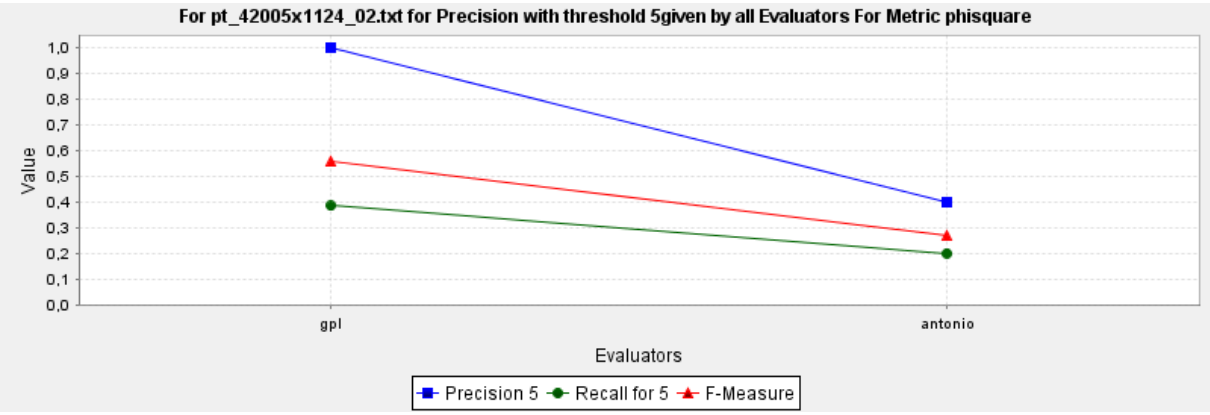


Figura 7.42 – Gráfico exemplificativo de relação de valores de precisão e cobertura para um documento e medida, para vários avaliadores.

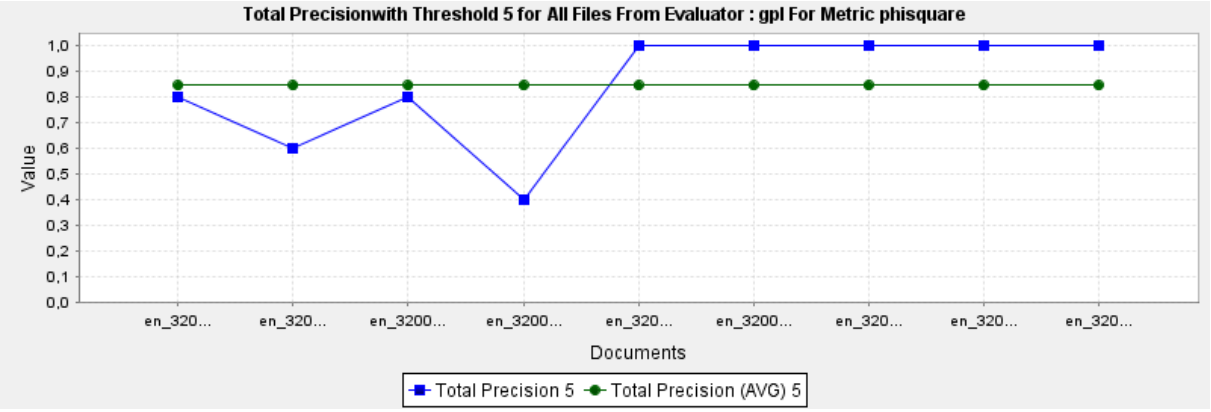


Figura 7.43- Gráfico que ilustra relação da precisão de cada documento com a média das precisões, para um avaliador e para uma dada medida

Threshold	Precision	Precision near good	Total Precision	Recall	F-Measure
5					
10					
15					
20					

Save Table Info

Figura 7.44 – Tabela onde serão apresentados os valores para a precisão, cobertura e f-measure

Threshold	Precision	Precision Near Good	Total Precision	Recall	F-Measure
5	0,200000000000000	0,000000000000000	0,200000000000000	0,111111111111111	0,142857142857143
10	0,222222222222222	0,000000000000000	0,222222222222222	0,222222222222222	0,222222222222222
15	0,142857142857143	0,000000000000000	0,142857142857143	0,222222222222222	0,173913043478261
20	0,263157894736842	0,000000000000000	0,263157894736842	0,555555555555556	0,357142857142857

Save Table Info

Figura 7.45 – Tabela onde serão apresentados os valores para a precisão, cobertura e f-measure populada.

Precision\Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Total Precision Avg (5)	0.727777777777778	0.638888888888889	0.462962962962963	0.424074074074074	0.622222222222222	0.516666666666667
Total Precision Avg (10)	0.725000000000000	0.660978835978836	0.355202821869488	0.353968253968254	0.613580246913580	0.483289241622575
Total Precision Avg (15)	0.680260480260480	0.640761090761091	0.347985347985348	0.351628001628002	0.620490620490620	0.453106153106153
Total Precision Avg (20)	0.621251385544471	0.645621201697053	0.345351327665569	0.334064941766180	0.626377422313955	0.414740896358543

Figura 7.46 - Tabela onde é apresentada a precisão total média, para todas as medidas avaliadas

Recall\Threshold	Phi^2	Least Tf-Ifd	Least M Rvar	Least M MI	Least M B Phi^2	Least M B Rvar
Recall Avg (5)	0.025265661790034	0.015737708490700	0.003319591048212	0.003485246365459	0.010652429700501	0.003815247998957
Recall Avg (10)	0.047349615947511	0.026737626482096	0.004549283154262	0.004489957714231	0.019398792387499	0.006081986080454
Recall Avg (15)	0.060678247978575	0.039228155236599	0.006016351940262	0.006099696044913	0.024145442689057	0.007682199298524
Recall Avg (20)	0.072873209929798	0.051940660635560	0.008621813958338	0.007891543758549	0.029198927886307	0.009130244977550

Figura 7.47 - Tabela onde é apresentada a cobertura média, para todas as medidas avaliadas

Nas seguintes figuras apresentamos a forma como calcular o valor da estatística Kappa para dois avaliadores, para um determinado documento e medida. Primeiro é necessário desbloquear a área de cálculo da estatística Kappa. Isso é alcançado fazendo a selecção da caixa de escolha presente na Figura 7.48. Esta acção fará com que o conteúdo da Figura 7.49 seja apresentada ao utilizador.

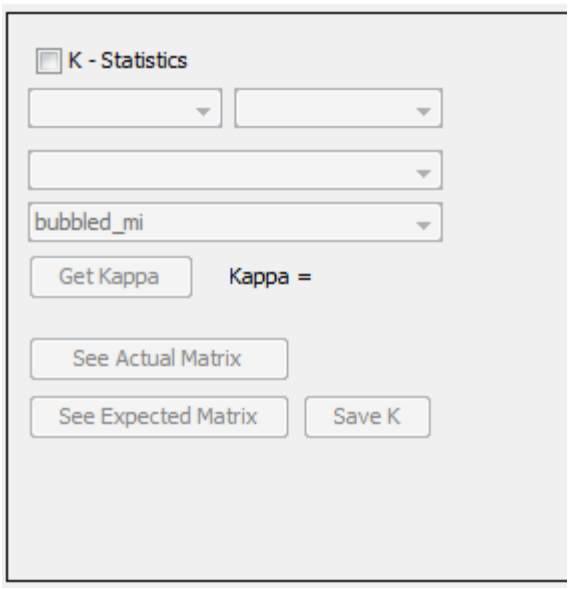


Figura 7.48 – Componente que permite o cálculo da estatística Kappa desactivada.

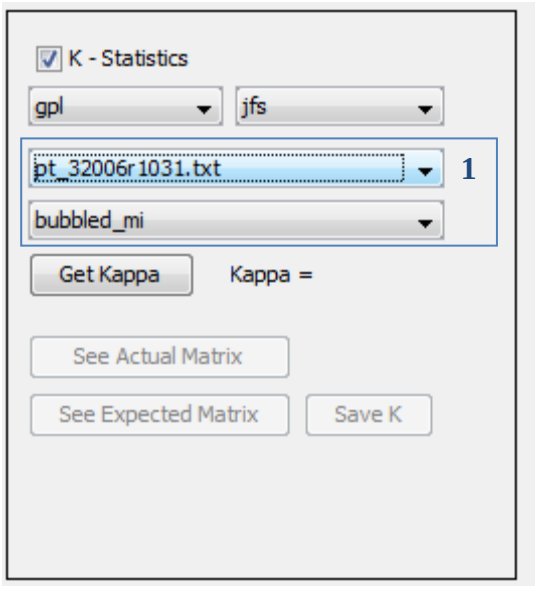


Figura 7.49 Componente que permite o cálculo da estatística Kappa activa

Na Figura 7.49 está identificado com (1) os componentes que permitem selecionar determinado ficheiro em comum entre os dois avaliadores, e uma medida, sobre a qual se queira ver o valor de Kappa. Para isso, o utilizador, após ter seleccionado o que pretende só tem de clicar no botão “Get Kappa” automaticamente verá o valor Kappa apresentado como se vê na Figura 7.50.

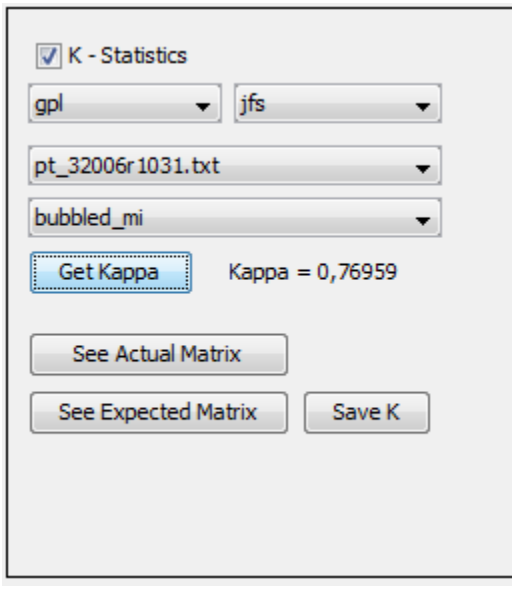


Figura 7.50 - – Componente que permite o cálculo da estatística Kappa com um exemplo.

Na Figura 7.50 também é possível observar que três botões foram desbloqueados quando se calculou o valor Kappa. Estes botões permitem ver as matrizes necessárias ao cálculo deste mesmo valor. Na Figura 7.51, podemos observar um exemplo de uma matriz confusão de resultados verificados. Já na Figura 7.52 podemos ver uma matriz confusão com resultados esperados. O Botão "Save Kappa" permite guardar em ficheiro a informação toda que foi necessária para calcular o valor Kappa.

	Good Descriptor	Near Good Descriptor	Bad Descriptor	Unkown	No Evaluation	Line Total
Good Descriptor	2	0	0	1	0	3
Near Good Descriptor	0	0	1	0	0	1
Bad Descriptor	0	0	1	0	0	1
Unkown	0	0	0	0	0	0
No Evaluation	0	0	0	0	20	20
Column Total	2	0	2	1	20	25

Save Table Info

Figura 7.51 – Matriz Confusão com resultados verificados entre dois avaliadores

	Good Descriptor	Near Good Descriptor	Bad Descriptor	Unkown	No Evaluation	Line Total
Good Descriptor	0.24	0.0	0.24	0.12	2.4	3.0
Near Good Descriptor	0.08	0.0	0.08	0.04	0.8	1.0
Bad Descriptor	0.08	0.0	0.08	0.04	0.8	1.0
Unkown	0.0	0.0	0.0	0.0	0.0	0.0
No Evaluation	1.6	0.0	1.6	0.8	16.0	20.0
Column Total	2.0	0.0	2.0	1.0	20.0	25.0

Save Table Info

Figura 7.52 - Matriz Confusão com resultados esperados entre dois avaliadores

Anexo 3 – Resultados

Neste anexo, serão apresentados tabelas e gráficos resultantes da análise das avaliações feitas por vários avaliadores a termos de vários documentos.

8.1 Cálculos da Estatística Kappa entre Prof. Joaquim Ferreira da Silva e o Prof. Gabriel Lopes para o documento pt_32006R0198.html

8.1.1 Kappa para a Medida Phi-Square

Este cálculo refere-se à medida *Phi-Square* para o documento pt_32006R0198.html[57].

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	9	0	1	1	0	11
	Near Good Descriptor	0	0	5	0	0	5
	Bad Descriptor	0	0	9	0	0	9
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	9	0	15	1	0	25

Tabela 8.1- Matriz Confusão de Resultados Verificados para Phi-Square

[57] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	3,960	0,000	6,600	0,440	0,000	10,900
	Near Good Descriptor	1,800	0,000	3,000	0,200	0,000	5,000
	Bad Descriptor	3,240	0,000	5,400	0,360	0,000	9,000
	Unkown	0,000	0,000	0,000	0,000	0,000	0,000
	No Evaluation	0,000	0,000	0,000	0,000	0,000	0,000
	Column Total	9,000	0,000	15,000	1,000	0,000	25,000

Tabela 8.2 - Matriz Confusão de Resultados Esperados para Phi-Square

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,552429667519181, o que dá aproximadamente 55.2% de concordância.

8.1.2 Kappa para a Medida Least Tf-Idf

Este cálculo refere-se à medida *Least Tf-Idf* para o documento pt_32006R0198.html[58]. Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	13	0	2	0	0	15
	Near Good Descriptor	0	0	2	0	0	2
	Bad Descriptor	0	1	7	0	0	8
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	13	1	11	0	0	25

Tabela 8.3 - Matriz Confusão de Resultados Verificados para Least Tf-Idf

[58] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	7,8	0,6	6,6	0	0	15
	Near Good Descriptor	1,04	0,08	0,88	0	0	2
	Bad Descriptor	4,16	0,32	3,52	0	0	8
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	13	1	11	0	0	25

Tabela 8.4 - Matriz Confusão de Resultados Esperados para Least Tf-Idf

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,63235294117647, o que dá aproximadamente 63.24% de concordância.

8.1.3 Kappa para a Medida Least Median Rvar

Este cálculo refere-se à medida *Least Median Rvar* para o documento pt_32006R0198.html[59].

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	2	2	7	1	0	12
	Near Good Descriptor	0	0	6	0	0	6
	Bad Descriptor	0	0	7	0	0	7
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	2	2	20	1	0	25

Tabela 8.5 - Matriz Confusão de Resultados Verificados para Least Median Rvar

[59] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	0,96	0,96	9,6	0,48	0	12
	Near Good Descriptor	0,48	0,48	4,8	0,24	0	6
	Bad Descriptor	0,56	0,56	5,6	0,28	0	7
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	2	2	20	1	0	25

Tabela 8.6 - Matriz Confusão de Resultados Esperados para Least Median Rvar

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,10913140311804, o que dá aproximadamente 11% de concordância.

8.1.4 Kappa para a Medida Least Median MI

Este cálculo refere-se à medida *Least Median MI* para o documento pt_32006R0198.html[60].

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	2	2	9	1	0	14
	Near Good Descriptor	0	0	5	0	0	5
	Bad Descriptor	1	0	5	0	0	6
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	2	19	1	0	25

Tabela 8.7 5 - Matriz Confusão de Resultados Verificados para Least Median MI

[60] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	1,68	1,12	10,64	0,56	0	14
	Near Good Descriptor	0,6	0,4	3,8	0,2	0	5
	Bad Descriptor	0,72	0,48	4,56	0,24	0	6
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	2	19	1	0	25

Tabela 8.8 5 - Matriz Confusão de Resultados Esperados para Least Median Rvar

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,0196078431372549, o que dá aproximadamente 1.96% de concordância.

8.1.5 Kappa para a Medida Least Bubbled Median Phi-Square

Este cálculo refere-se à medida *Least Bubbled Median Phi-Square* para o documento pt_32006R0198.html[61].

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	8	0	1	0	0	9
	Near Good Descriptor	0	0	3	0	0	3
	Bad Descriptor	0	1	12	0	0	13
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	8	1	16	0	0	25

Tabela 8.9 - Matriz Confusão de Resultados Verificados para Least Bubbled Median Phi-Square

[61] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	2,88	0,36	5,76	0	0	9
	Near Good Descriptor	0,96	0,12	1,92	0	0	3
	Bad Descriptor	4,16	0,52	8,32	0	0	13
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	8	1	16	0	0	25

Tabela 8.10 - Matriz Confusão de Resultados Esperados para Least Bubbled Median Phi-Square

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,634502923976608, o que dá aproximadamente 63.5% de concordância.

8.1.6 Kappa para a Medida Least Bubbled Median Rvar

Este cálculo refere-se à medida *Least Bubbled Median Rvar* para o documento pt_32006R0198.html[62].

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	5	1	7	3	0	16
	Near Good Descriptor	0	0	3	0	0	3
	Bad Descriptor	0	0	6	0	0	6
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	5	1	16	3	0	25

Tabela 8.11 - Matriz Confusão de Resultados Verificados para Least Bubbled Median Rvar

[62] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	3,2	0,64	10,24	1,92	0	16
	Near Good Descriptor	0,6	0,12	1,92	0,36	0	3
	Bad Descriptor	1,2	0,24	3,84	0,72	0	6
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	5	1	16	3	0	25

Tabela 8.12 - Matriz Confusão de Resultados Esperados para Least Bubbled Median Rvar

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,2152466367713, o que dá aproximadamente 21.52% de concordância.

8.2 Lista de Termos Avaliados pelo Avaliador Prof. Gabriel Lopes para o documento pt_32006R0198.html

Apresenta-se de seguida a listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes, para as medidas pedidas.

8.2.1 PhiSquare

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
formação profissional contínua	0,008977472052384	good topic descriptor
profissional contínua	0,008977472052384	bad descriptor
contínua	0,008257084363260	bad descriptor
formação profissional	0,007613838869853	good topic descriptor
profissional	0,006731434220435	bad descriptor
em horas	0,005207533750025	bad descriptor
cursos de formação profissional contínua	0,005096688636165	good topic descriptor
cursos	0,005080076295244	good topic descriptor
cursos de formação	0,005064663891633	good topic descriptor
formação	0,004140313788898	good topic descriptor
nenhum valor em falta	0,003545069493752	bad descriptor
valor em falta	0,003545069493752	bad descriptor
nenhum valor	0,003545069493752	bad descriptor
número	0,003345129880868	bad descriptor
número total	0,003309304724491	bad descriptor
imputação	**0,002547809415785**	**unkonwn**
profissional inicial	0,002534794484038	bad descriptor
tempo de trabalho	0,002437012852767	good topic descriptor
remunerado	0,002437012852767	bad descriptor
nenhum	0,002421652204649	bad descriptor
empresas	0,002204694848287	good topic descriptor
amostragem	0,002200631608461	good topic descriptor
inicial	0,002125444852977	bad descriptor
empregadas	0,002120291214962	bad descriptor
— sem classificação	0,001883060370466	bad descriptor

Tabela 8.13 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Phi-Square

8.2.2 Least Tf-Idf

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
profissional	0,017270167990526	bad descriptor
contínua	0,016727894319951	bad descriptor
profissional contínua	0,016727894319951	bad descriptor
cursos de formação profissional contínua	0,012184515615767	good topic descriptor
cursos	0,012184515615767	good topic descriptor
formação profissional contínua	0,009593030169595	good topic descriptor
formação	0,009593030169595	good topic descriptor
cursos de formação	0,009593030169595	good topic descriptor
formação profissional	0,009593030169595	good topic descriptor
cursos internos de formação	0,009593030169595	good topic descriptor
imputação	**0,009187329625273**	**unkonwn**
formação específicas das pessoas empregadas	0,009174378153781	near good descriptor
contínua para pessoas empregadas	0,009174378153781	bad descriptor
empregadas	0,009174378153781	bad descriptor
empresas	0,008973854651220	good topic descriptor
empregadas em empresas	0,008973854651220	bad descriptor
profissional nas empresas	0,008973854651220	bad descriptor
formação profissional nas empresas	0,008973854651220	good topic descriptor
empresas que fazem formação	0,008973854651220	good topic descriptor
remunerado para cursos	0,008787880511131	bad descriptor
remunerado	0,008787880511131	bad descriptor
remunerado em cursos	0,008787880511131	bad descriptor
participantes em cursos	0,006961567700693	good topic descriptor
participantes	0,006961567700693	good topic descriptor
participantes em formação profissional	0,006961567700693	good topic descriptor

Tabela 8.14 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Least Tf-Idf

8.2.3 Least Median Rvar

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
estatísticas-chave	17,999999999999996	good topic descriptor
significativamente	17,999999999999996	bad descriptor
pormenorizadamente	17,999999999999996	bad descriptor
subpopulações-alvo	17,999999999999996	near good descriptor
electronicamente	15,999999999999996	bad descriptor
horvitz-thompson	15,999999999999996	good topic descriptor
socioeconómicas	14,999999999999996	bad descriptor
variáveis-chave	14,999999999999996	good topic descriptor
variável-chave	14,000000000000000	good topic descriptor
estratificados	13,999999999999996	bad descriptor
probabilística	13,999999999999996	near good descriptor
corresponderam	13,999999999999996	bad descriptor
pormenorizados	13,999999999999996	bad descriptor
população-alvo	13,999999999999996	near good descriptor
sobrecobertura	13,999999999999996	near good descriptor
significativamente melhorados	13,999999999999996	bad descriptor
probabilística estratificada	13,499999999999996	bad descriptor
variável-base	13,000000000000000	good topic descriptor
empresas-mães	12,999999999999996	bad descriptor
laboratoriais	12,999999999999996	bad descriptor
preenchimento	12,999999999999996	bad descriptor
destacamentos	**12,999999999999996**	**unkonwn**
identificadas	12,999999999999996	bad descriptor
não-respostas	12,999999999999996	good topic descriptor
problemáticas	12,999999999999996	bad descriptor

Tabela 8.15 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Least Median Rvar

8.2.4 Least Median MI

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
estatísticas-chave	46,359290347154630	good topic descriptor
significativamente	46,359290347154630	bad descriptor
pormenorizadamente	46,359290347154630	bad descriptor
subpopulações-alvo	46,359290347154630	near good descriptor
electronicamente	41,208258086359670	bad descriptor
horvitz-thompson	41,208258086359670	good topic descriptor
socioeconómicas	38,632741955962190	bad descriptor
variáveis-chave	38,632741955962190	good topic descriptor
estratificados	36,057225825564714	bad descriptor
probabilística	36,057225825564714	near good descriptor
corresponderam	36,057225825564714	bad descriptor
pormenorizados	36,057225825564714	bad descriptor
variável-chave	36,057225825564714	good topic descriptor
população-alvo	36,057225825564714	near good descriptor
sobrecobertura	36,057225825564714	near good descriptor
significativamente melhorados	36,057225825564714	bad descriptor
probabilística estratificada	34,769467760365970	bad descriptor
empresas-mães	33,481709695167230	bad descriptor
laboratoriais	33,481709695167230	bad descriptor
preenchimento	33,481709695167230	bad descriptor
destacamentos	**33,481709695167230**	**unkonwn**
identificadas	33,481709695167230	bad descriptor
não-respostas	33,481709695167230	good topic descriptor
problemáticas	33,481709695167230	bad descriptor
questionários	33,481709695167230	good topic descriptor

Tabela 8.16 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Least Median MI

8.2.5 Least Bubbled Median Phi-Square

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
contínua	0,062639410875556	bad descriptor
profissional	0,056544502411978	bad descriptor
profissional contínua	0,047120418676649	bad descriptor
formação profissional	0,041244206779647	good topic descriptor
empresas-mães	0,040936954447726	bad descriptor
cursos de formação profissional contínua	0,040640610361951	good topic descriptor
amostragem	0,038514649217131	good topic descriptor
amostrais	0,034663184295418	good topic descriptor
empresarial	0,034638961455768	bad descriptor
formação profissional contínua	0,032995365423718	good topic descriptor
formação	0,032995365423718	good topic descriptor
variáveis-chave	0,032924777689455	good topic descriptor
amostragem incluídas na amostra	0,030811719373705	bad descriptor
variável-chave	0,030729792510158	good topic descriptor
cursos	0,030480457771463	good topic descriptor
cursos internos de formação	0,028870944745753	good topic descriptor
variável-base	0,028534807330861	good topic descriptor
formação no desempenho empresarial	0,028340968463810	good topic descriptor
imputações	**0,027694086088190**	**unkonwn**
amostra	0,026960254451992	good topic descriptor
empresas nos estratos de amostragem	0,025191971967832	good topic descriptor
empresas	0,025191971967832	good topic descriptor
profissional nas empresas	0,025191971967832	bad descriptor
formação profissional nas empresas	0,025191971967832	good topic descriptor
formação profissional contínua da empresa	0,025191971967832	good topic descriptor

Tabela 8.17 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Least Bubbled Median Phi-Square

8.2.6 Least Bubbled Median Rvar

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
subpopulações-alvo	17,999999999999996	near good descriptor
horvitz-thompson	15,999999999999996	good topic descriptor
não-respostas	13,000000000000004	good topic descriptor
destacamentos	**12,999999999999996**	**unkonwn**
influenciaram	12,999999999999996	bad descriptor
não-resposta	12,000000000000004	good topic descriptor
reponderação	11,999999999999996	bad descriptor
não-formação	11,999999999999996	good topic descriptor
pac=c3tot*a5	11,999999999999996	bad descriptor
coeficientes	**11,999999999999996**	**unkonwn**
subcobertura	11,999999999999996	near good descriptor
planificação	11,999999999999996	bad descriptor
acessibilidade	11,943045311153242	bad descriptor
comentários	11,000000000000002	bad descriptor
coeficiente	**10,999999999999998**	**unkonwn**
codificação	10,999999999999998	bad descriptor
sobrecobertura	10,842529794442926	near good descriptor
probabilística	10,383412029287300	near good descriptor
ventilação	10,000000000000002	bad descriptor
honorários	10,000000000000002	bad descriptor
calcula-se	10,000000000000000	bad descriptor
imputações	**10,000000000000000**	**unkonwn**
calcularão	10,000000000000000	bad descriptor
subamostra	9,999999999999998	good topic descriptor
recalcular	9,999999999999998	bad descriptor

Tabela 8.18 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento pt32006R198.html na medida Least Bubbled Median Rvar

8.3 Lista de Termos Avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva para o documento pt_32006R0198.html

Apresenta-se de seguida listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva, para as medidas pedidas.

8.3.1 Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
formação profissional contínua	0,008977472052384	good topic descriptor
profissional contínua	0,008977472052384	good topic descriptor
contínua	0,008257084363260	bad descriptor
formação profissional	0,007613838869853	good topic descriptor
profissional	0,006731434220435	near good descriptor
em horas	0,005207533750025	bad descriptor
cursos de formação profissional contínua	0,005096688636165	good topic descriptor
cursos	0,005080076295244	good topic descriptor
cursos de formação	0,005064663891633	good topic descriptor
formação	0,004140313788898	good topic descriptor
nenhum valor em falta	0,003545069493752	bad descriptor
valor em falta	0,003545069493752	near good descriptor
nenhum valor	0,003545069493752	bad descriptor
número	0,003345129880868	near good descriptor
número total	0,003309304724491	near good descriptor
imputação	0,002547809415785	good topic descriptor
profissional inicial	0,002534794484038	bad descriptor
tempo de trabalho	0,002437012852767	good topic descriptor
remunerado	0,002437012852767	bad descriptor
nenhum	0,002421652204649	bad descriptor
empresas	0,002204694848287	good topic descriptor
amostragem	0,002200631608461	good topic descriptor
inicial	0,002125444852977	bad descriptor
empregadas	0,002120291214962	near good descriptor
— sem classificação	0,001883060370466	bad descriptor

Tabela 8.19 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Phi-Square

8.3.2 Least Tf-Idf

Termos	Valor da medida	Avaliação dada ao termo pelo Avaliador
profissional	0,017270167990526	near good descriptor
contínua	0,016727894319951	bad descriptor
profissional contínua	0,016727894319951	good topic descriptor
cursos de formação profissional contínua	0,012184515615767	good topic descriptor
cursos	0,012184515615767	good topic descriptor
formação profissional contínua	0,009593030169595	good topic descriptor
formação	0,009593030169595	good topic descriptor
cursos de formação	0,009593030169595	good topic descriptor
formação profissional	0,009593030169595	good topic descriptor
cursos internos de formação	0,009593030169595	good topic descriptor
imputação	0,009187329625273	good topic descriptor
formação específicas das pessoas empregadas	0,009174378153781	good topic descriptor
contínua para pessoas empregadas	0,009174378153781	bad descriptor
empregadas	0,009174378153781	near good descriptor
empresas	0,008973854651220	good topic descriptor
empregadas em empresas	0,008973854651220	near good descriptor
profissional nas empresas	0,008973854651220	bad descriptor
formação profissional nas empresas	0,008973854651220	good topic descriptor
empresas que fazem formação	0,008973854651220	near good descriptor
remunerado para cursos	0,008787880511131	bad descriptor
remunerado	0,008787880511131	bad descriptor
remunerado em cursos	0,008787880511131	bad descriptor
participantes em cursos	0,006961567700693	good topic descriptor
participantes	0,006961567700693	good topic descriptor
participantes em formação profissional	0,006961567700693	good topic descriptor

Tabela 8.20 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Least Tf-Idf

8.3.3 Least Median Rvar

Termos	Valores da medida	Avaliação dada ao termo pelo Avaliador
estatísticas-chave	17,999999999999996	good topic descriptor
significativamente	17,999999999999996	bad descriptor
pormenorizadamente	17,999999999999996	bad descriptor
subpopulações-alvo	17,999999999999996	good topic descriptor
electronicamente	15,999999999999996	bad descriptor
horvitz-thompson	15,999999999999996	good topic descriptor
socioeconómicas	14,999999999999996	near good descriptor
variáveis-chave	14,999999999999996	good topic descriptor
variável-chave	14,000000000000000	good topic descriptor
estratificados	13,999999999999996	near good descriptor
probabilística	13,999999999999996	near good descriptor
corresponderam	13,999999999999996	bad descriptor
pormenorizados	13,999999999999996	bad descriptor
população-alvo	13,999999999999996	good topic descriptor
sobrecobertura	13,999999999999996	good topic descriptor
significativamente melhorados	13,999999999999996	near good descriptor
probabilística estratificada	13,499999999999996	near good descriptor
variável-base	13,000000000000000	good topic descriptor
empresas-mães	12,999999999999996	good topic descriptor
laboratoriais	12,999999999999996	bad descriptor
preenchimento	12,999999999999996	good topic descriptor
destacamentos	12,999999999999996	good topic descriptor
identificadas	12,999999999999996	bad descriptor
não-respostas	12,999999999999996	good topic descriptor
problemáticas	12,999999999999996	near good descriptor

Tabela 8.21 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Least Median Rvar

8.3.4 Least Median MI

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
estatísticas-chave	46,359290347154630	good topic descriptor
significativamente	46,359290347154630	bad descriptor
pormenorizadamente	46,359290347154630	bad descriptor
subpopulações-alvo	46,359290347154630	good topic descriptor
electronicamente	41,208258086359670	bad descriptor
horvitz-thompson	41,208258086359670	good topic descriptor
socioeconómicas	38,632741955962190	near good descriptor
variáveis-chave	38,632741955962190	good topic descriptor
estratificados	36,057225825564714	near good descriptor
probabilística	36,057225825564714	near good descriptor
corresponderam	36,057225825564714	bad descriptor
pormenorizados	36,057225825564714	bad descriptor
variável-chave	36,057225825564714	good topic descriptor
população-alvo	36,057225825564714	good topic descriptor
sobrecobertura	36,057225825564714	good topic descriptor
significativamente melhorados	36,057225825564714	near good descriptor
probabilística estratificada	34,769467760365970	near good descriptor
empresas-mães	33,481709695167230	good topic descriptor
laboratoriais	33,481709695167230	bad descriptor
preenchimento	33,481709695167230	good topic descriptor
destacamentos	33,481709695167230	good topic descriptor
identificadas	33,481709695167230	bad descriptor
não-respostas	33,481709695167230	good topic descriptor
problemáticas	33,481709695167230	near good descriptor
questionários	33,481709695167230	good topic descriptor

Tabela 8.22 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Least Median MI

8.3.5 Least Bubbled Median Phi-Square

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
contínua	0,062639410875556	bad descriptor
profissional	0,056544502411978	near good descriptor
profissional contínua	0,047120418676649	good topic descriptor
formação profissional	0,041244206779647	good topic descriptor
empresas-mães	0,040936954447726	good topic descriptor
cursos de formação profissional contínua	0,040640610361951	good topic descriptor
amostragem	0,038514649217131	good topic descriptor
amostrais	0,034663184295418	near good descriptor
empresarial	0,034638961455768	near good descriptor
formação profissional contínua	0,032995365423718	good topic descriptor
formação	0,032995365423718	good topic descriptor
variáveis-chave	0,032924777689455	good topic descriptor
amostragem incluídas na amostra	0,030811719373705	good topic descriptor
variável-chave	0,030729792510158	good topic descriptor
cursos	0,030480457771463	good topic descriptor
cursos internos de formação	0,028870944745753	good topic descriptor
variável-base	0,028534807330861	good topic descriptor
formação no desempenho empresarial	0,028340968463810	good topic descriptor
imputações	0,027694086088190	good topic descriptor
amostra	0,026960254451992	good topic descriptor
empresas nos estratos de amostragem	0,025191971967832	good topic descriptor
empresas	0,025191971967832	good topic descriptor
profissional nas empresas	0,025191971967832	bad descriptor
formação profissional nas empresas	0,025191971967832	good topic descriptor
formação profissional contínua da empresa	0,025191971967832	good topic descriptor

Tabela 8.23 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Least Bubbled Median Phi-Square

8.3.6 Least Bubbled Median Rvar

Termos	Valores da Medida	Avaliação dada ao termo pelo Avaliador
subpopulações-alvo	17,999999999999996	good topic descriptor
horvitz-thompson	15,999999999999996	good topic descriptor
não-respostas	13,000000000000004	good topic descriptor
destacamentos	12,999999999999996	good topic descriptor
influenciaram	12,999999999999996	bad descriptor
não-resposta	12,000000000000004	good topic descriptor
reponderação	11,999999999999996	good topic descriptor
não-formação	11,999999999999996	good topic descriptor
pac=c3tot*a5	11,999999999999996	bad descriptor
coeficientes	11,999999999999996	good topic descriptor
subcobertura	11,999999999999996	good topic descriptor
planificação	11,999999999999996	good topic descriptor
acessibilidade	11,943045311153242	good topic descriptor
comentários	11,000000000000002	good topic descriptor
coeficiente	10,999999999999998	good topic descriptor
codificação	10,999999999999998	good topic descriptor
sobrecobertura	10,842529794442926	good topic descriptor
probabilística	10,383412029287300	near good descriptor
ventilação	10,000000000000002	good topic descriptor
honorários	10,000000000000002	good topic descriptor
calcula-se	10,000000000000000	bad descriptor
imputações	10,000000000000000	good topic descriptor
calcularão	10,000000000000000	bad descriptor
subamostra	9,999999999999998	good topic descriptor
recalcular	9,999999999999998	bad descriptor

Tabela 8.24 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento pt32006R198.html na medida Least Bubbled Median Rvar

8.4 Lista de Termos Apresentados aos Avaliadores para outras medidas

8.4.1 Rvar

Termos	Valor da Medida
totais da formação profissional inicial	1,00
nace_sp e por size_sp	1,00
desvios	1,00
mão-de-obra relativos	1,00
ponderação	1,00
ventilação	1,00
métodos utilizados	1,00
ponderações	1,00
não-resposta por unidade	1,00
estrato definido	1,00
alfanum	1,00
cada um dos campos nace	1,00
função da nace	1,00
comprimento	1,00
nace e do grupo	1,00
custos dos cursos	1,00
formais	1,00
resposta por unidade	1,00
questionário	1,00
relativamente aos seguintes pontos	1,00
prestaram	1,00
custos dos cursos de formação	1,00
campos nace	1,00
variável extra	1,00
papel da estrutura	1,00

Tabela 8.25 - Lista de Termos para a medida Rvar para o ficheiro pt_32006R0198.html

8.4.2 MI

Termos	Valor da Medida
estratificados	2,57551613
imputações não serão permitidas	2,57551613
se recomenda	2,57551613
qualificações formais — empregadas	2,57551613
totais da formação profissional inicial	2,57551613
software de avaliação da variância	2,57551613
empresas-mães	2,57551613
— com idade igual	2,57551613
preferiu conceder	2,57551613
gestão e administração	2,57551613
actuais e potenciais	2,57551613
permitidas se mais	2,57551613
variáveis identificadas	2,57551613
deve ser estabelecido o primeiro	2,57551613
identificadas no anexo i	2,57551613
socioeconómicas	2,57551613
registadas como	2,57551613
concretizadas em valores em falta	2,57551613
empresas-mães / associadas	2,57551613
partilhado	2,57551613
inquérito e outro inquérito	2,57551613
pessoas contratadas	2,57551613
empresas e à estrutura	2,57551613
demasiado elevados para a empresa	2,57551613
ventilação de correcções	2,57551613

Tabela 8.26 - Lista de Termos para a medida MI para o ficheiro pt_32006R0198.html

8.4.3 Tf-Idf

Termo	Valor da Medida
formação profissional contínua	0,0323554
profissional contínua	0,0323554
em horas	0,0187741
cursos de formação profissional contínua	0,0183747
profissional	0,0172702
formação profissional	0,0168196
contínua	0,0167279
cursos de formação	0,0131199
nenhum valor em falta	0,0127824
valor em falta	0,0127824
nenhum valor	0,0127824
cursos	0,0121845
número total	0,0098071
formação	0,009593
imputação	0,0091873
empregadas	0,0091744
empresas	0,0089739
tempo de trabalho	0,0087879
remunerado	0,0087879
profissional inicial	0,0075926
pessoas empregadas	0,0072762
participantes	0,0069616
— sem classificação	0,0067906
nenhum	0,0066086
remunerado —	0,0063912

Tabela 8.27 - Lista de Termos para a medida Tf-Idf para o ficheiro pt_32006R0198.html

8.5 Gráficos das Precisões para o Avaliador Prof. Gabriel Lopes para o documento pt_32006R0198.html

As seguintes figuras apresentam os gráficos com as precisões, cobertura e F-Measure, considerados mais demonstrativos e que foram obtidas da análise dos resultados do avaliador Prof. Gabriel Lopes para o documento pt_32006R0198.html[63]. Os gráficos mostram os valores de precisão para 5, 10, 15 e 20.

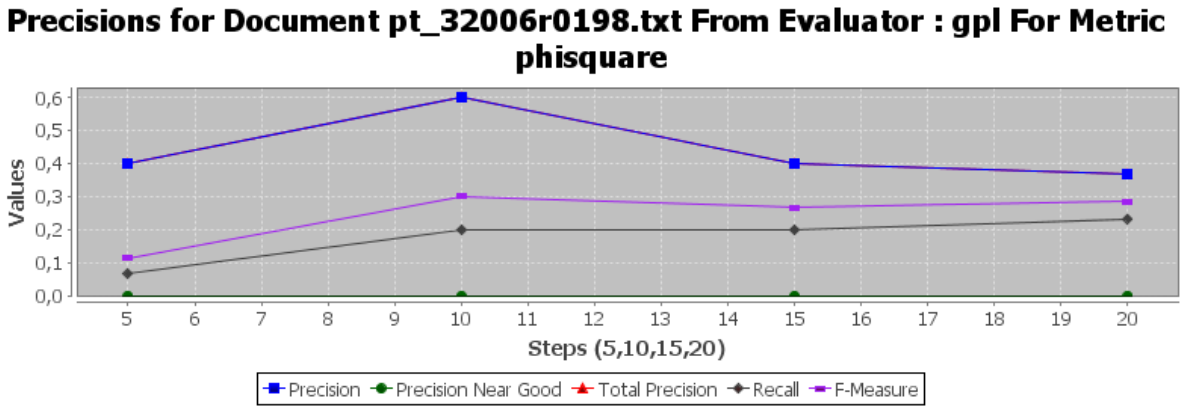


Figura 8.1 - Valores de Precisão, Cobertura e F-Measure para Phi-Square

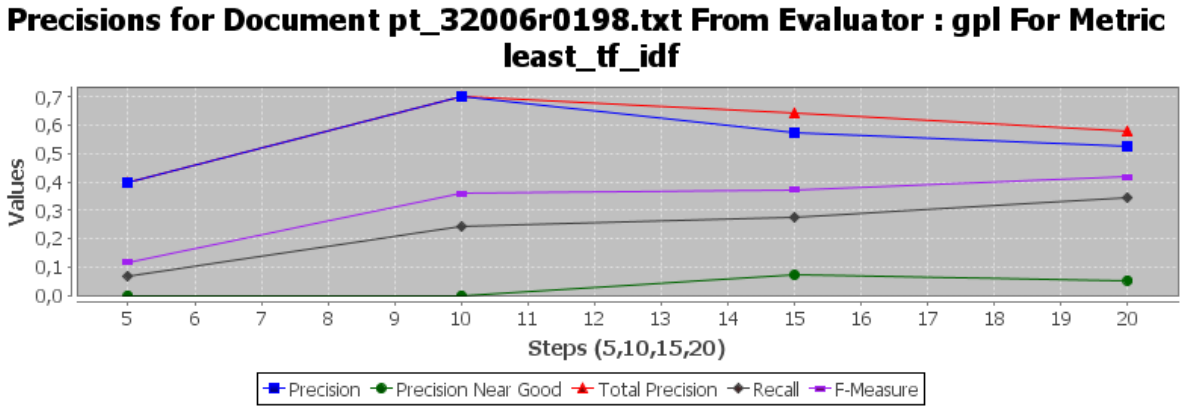


Figura 8.2 - Valores de Precisão, Cobertura e F-Measure para Least Tf-Idf

[63] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

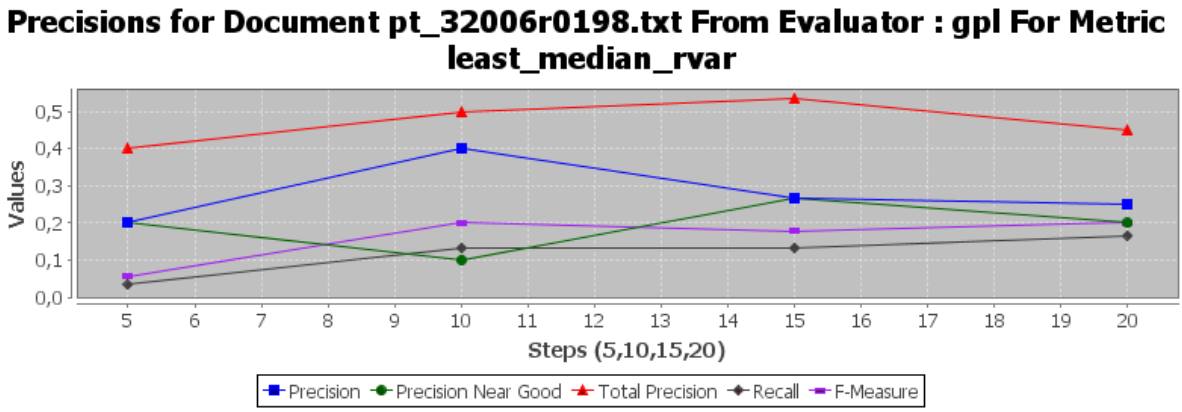


Figura 8.3 - Valores de Precisão, Cobertura e F-Measure para Least Median Rvar

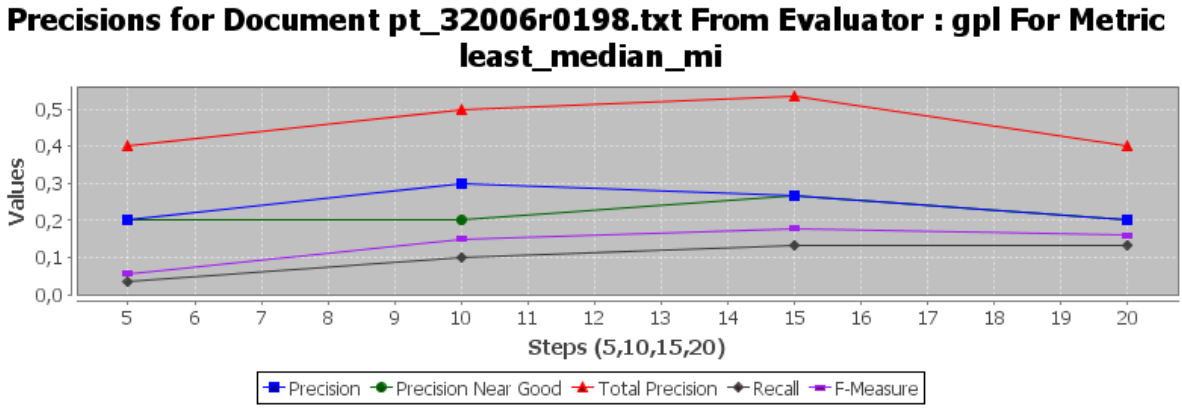


Figura 8.4 - Valores de Precisão, Cobertura e F-Measure para Least Median MI

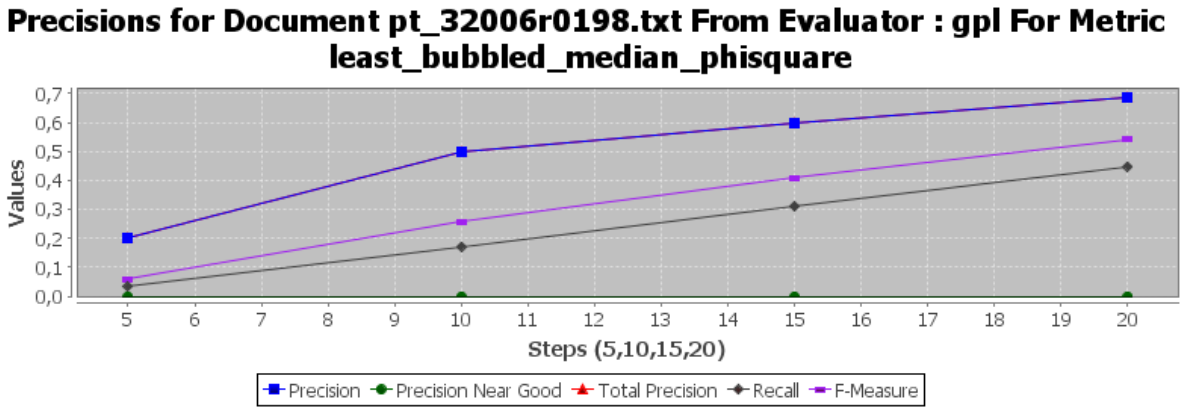


Figura 8.5 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Phi-Square

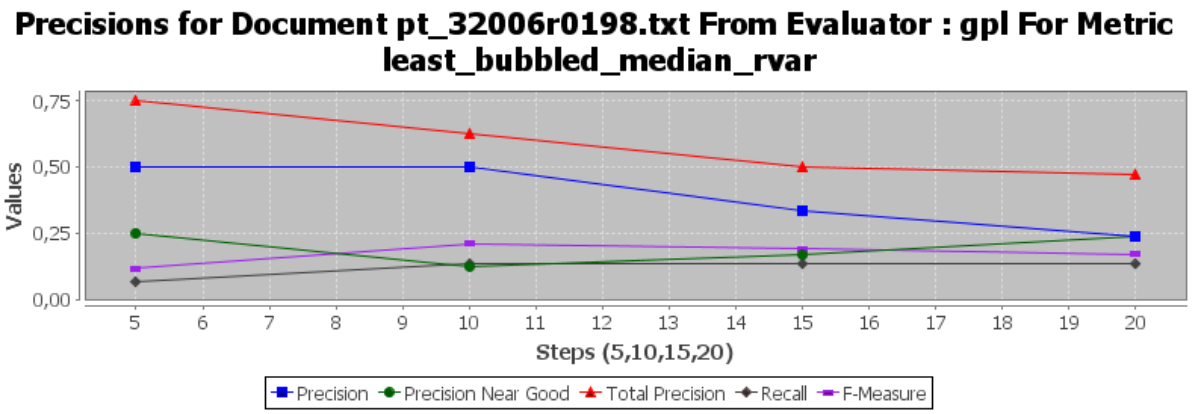


Figura 8.6 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Rvar

8.6 Gráficos da Precisão Total para todos os documentos em português avaliados pelo Avaliador Prof. Gabriel Lopes

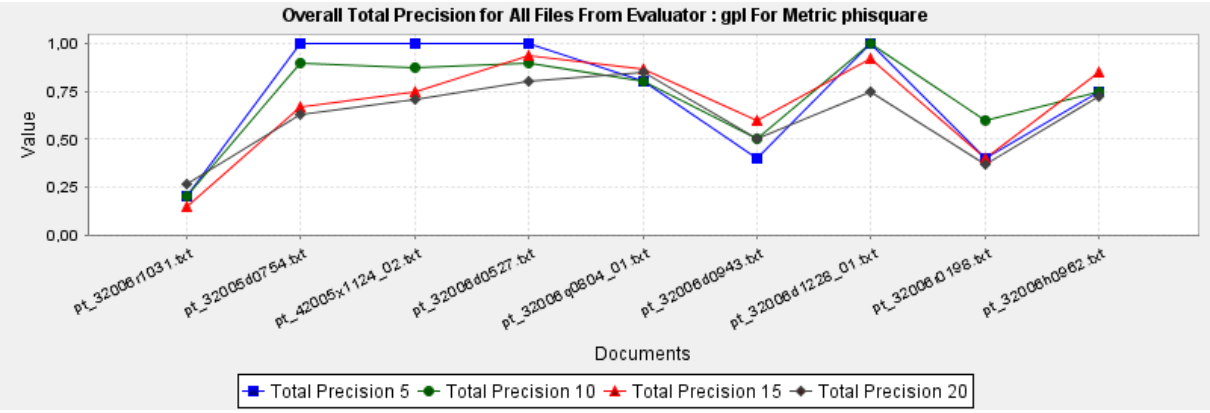


Figura 8.7 - Precisão total para todos os documentos, para a medida Phi-Square

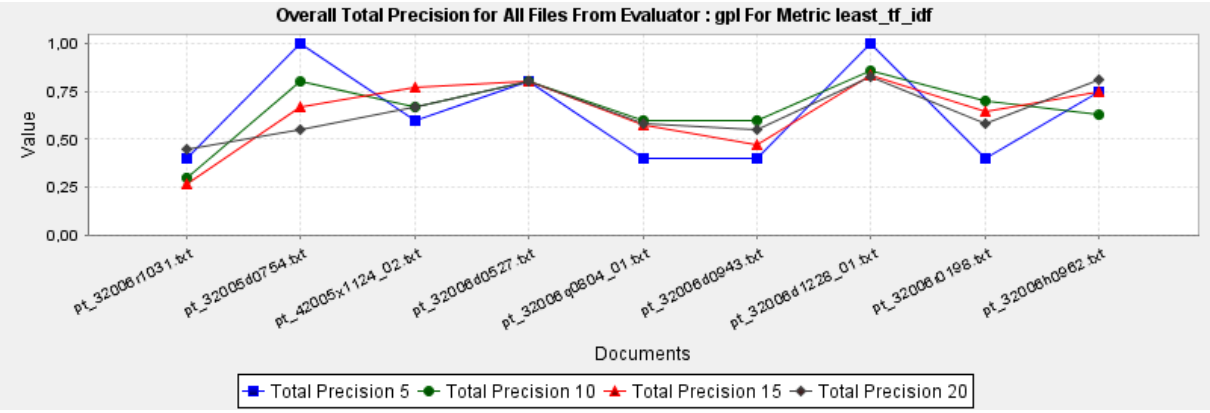


Figura 8.8 - Precisão total para todos os documentos, para a medida Least Tf-Idf

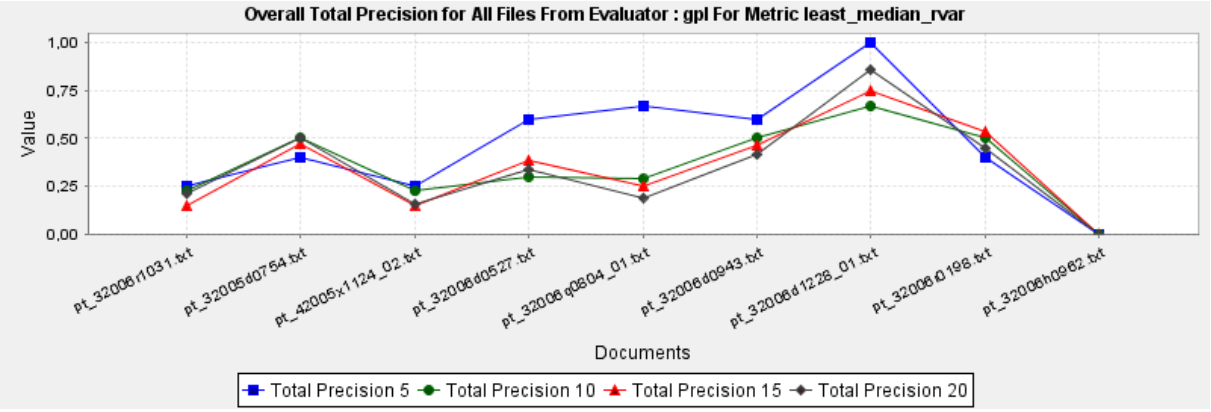


Figura 8.9 - Precisão total para todos os documentos em Português, para a medida Least Median Rvar

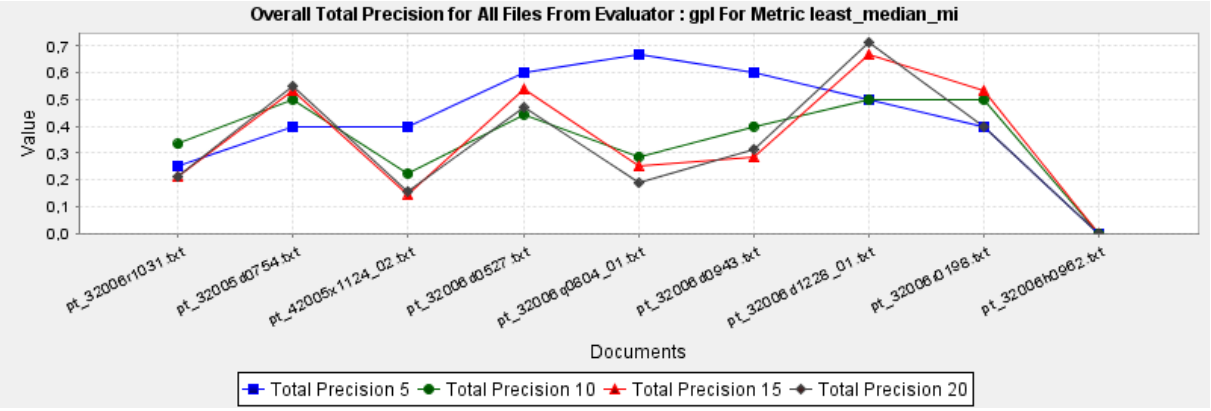


Figura 8.10 - Precisão total para todos os documentos em Português, para a medida Least Median MI

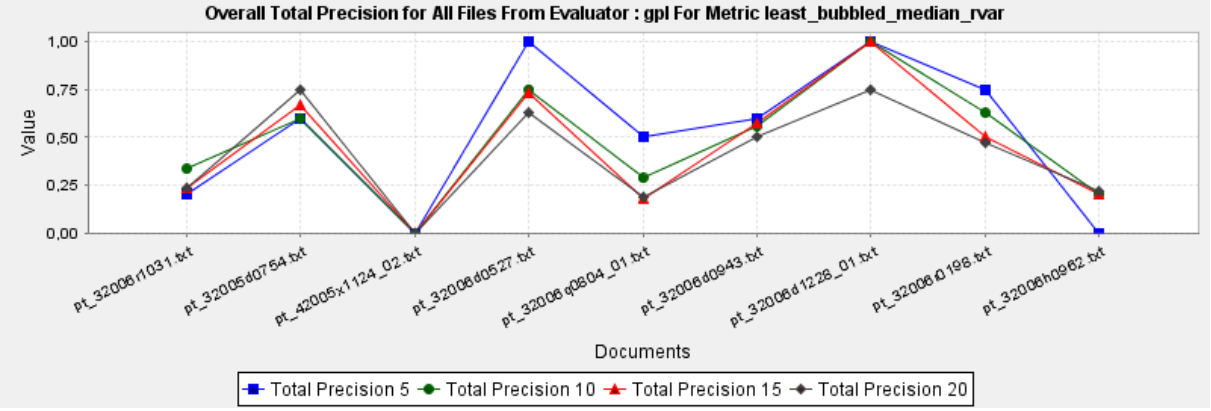


Figura 8.11 - Precisão total para todos os documentos em Português, para a medida Least Bubbled Median Phi-Square

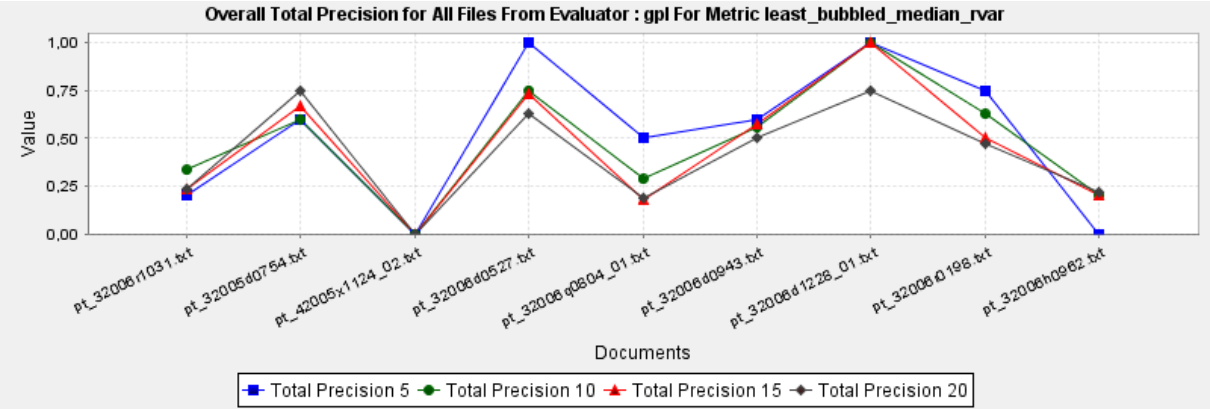


Figura 8.12 - Precisão total para todos os documentos em Português, para a medida Least Bubbled Median Rvar

8.7 Gráficos da Precisão Total versus Média da Precisão Total para todos os documentos em português avaliados pelo Avaliador Prof. Gabriel Lopes

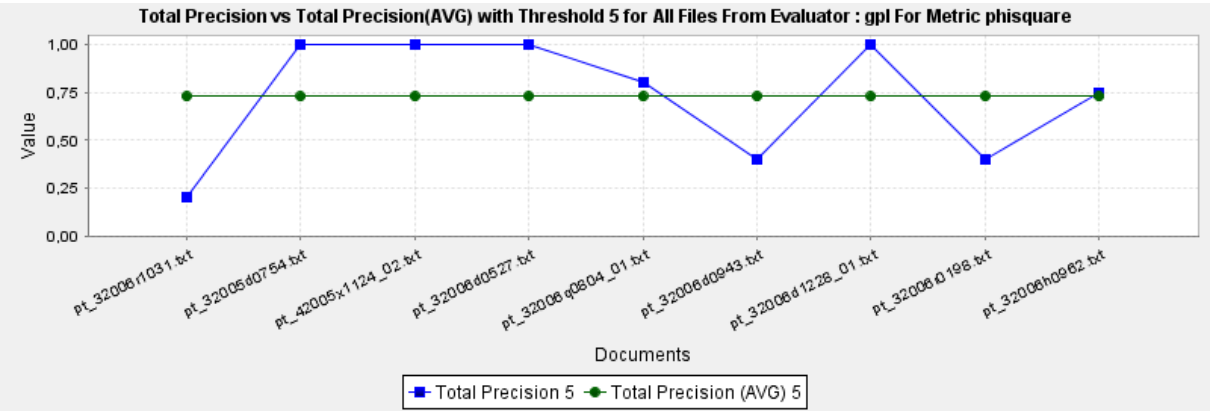


Figura 8.13 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 5

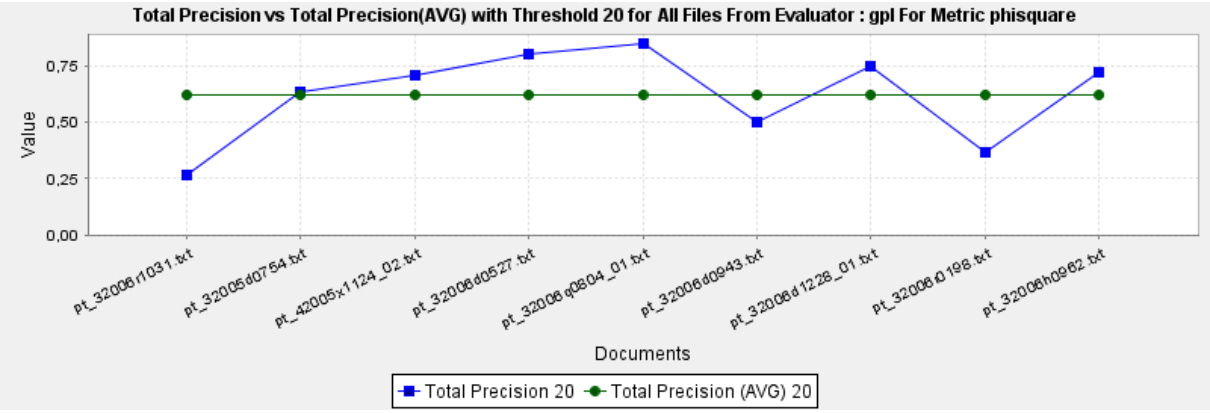


Figura 8.14 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 20

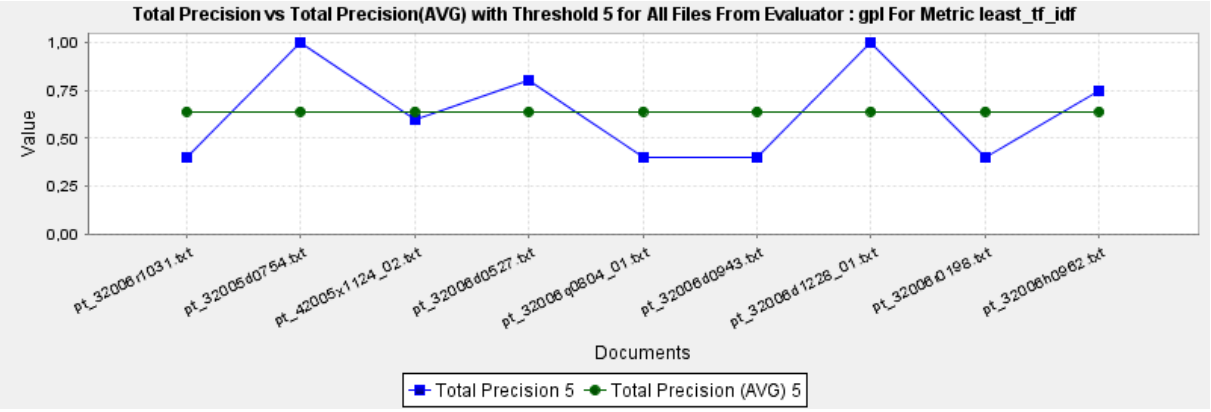


Figura 8.15 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Tf-Idf, com o limite 5

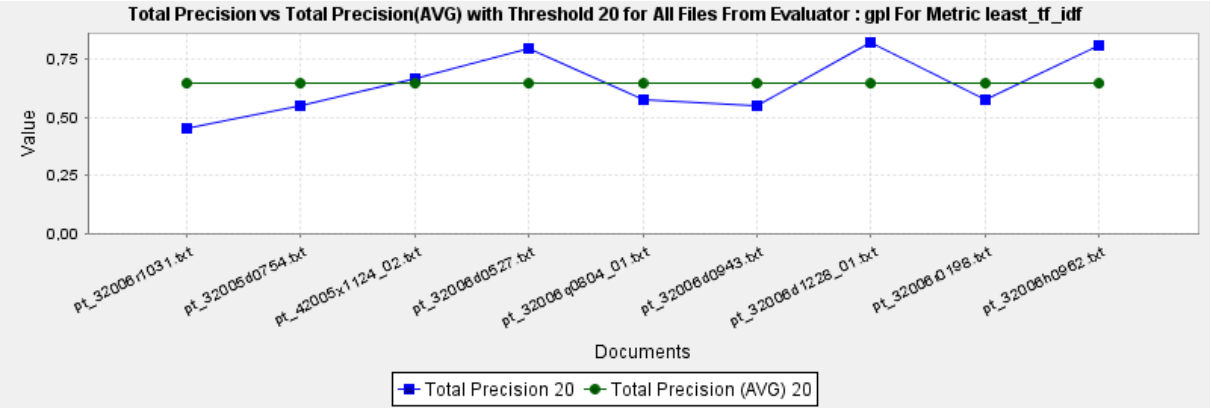


Figura 8.16 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Tf-Idf, com o limite 20

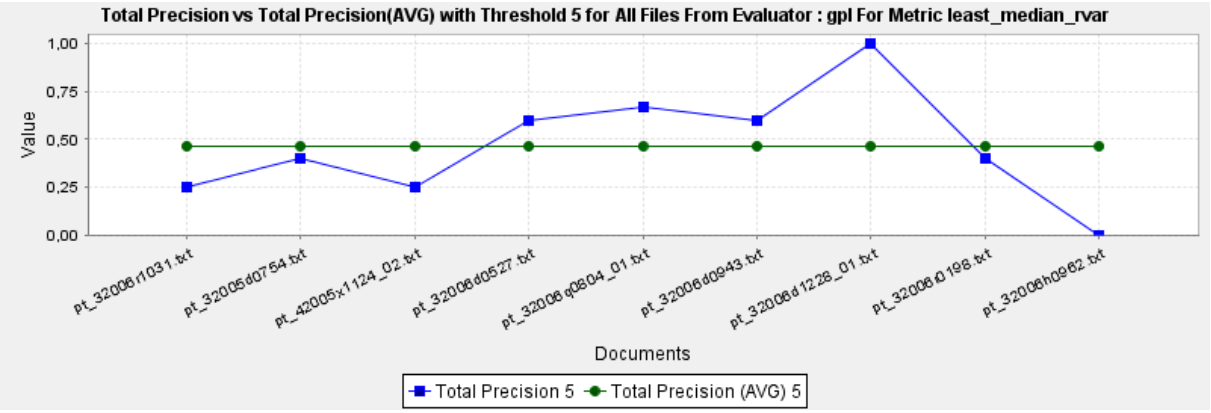


Figura 8.17 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 5

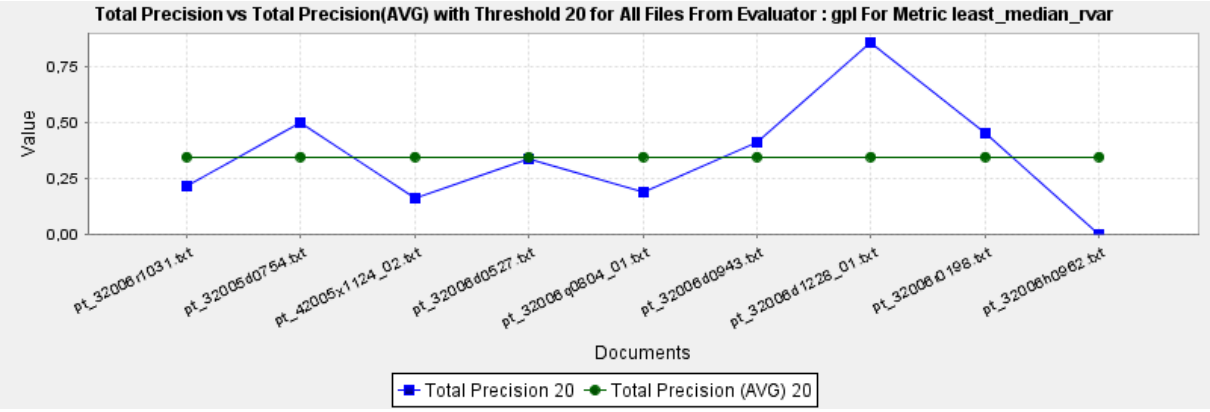


Figura 8.18 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 20

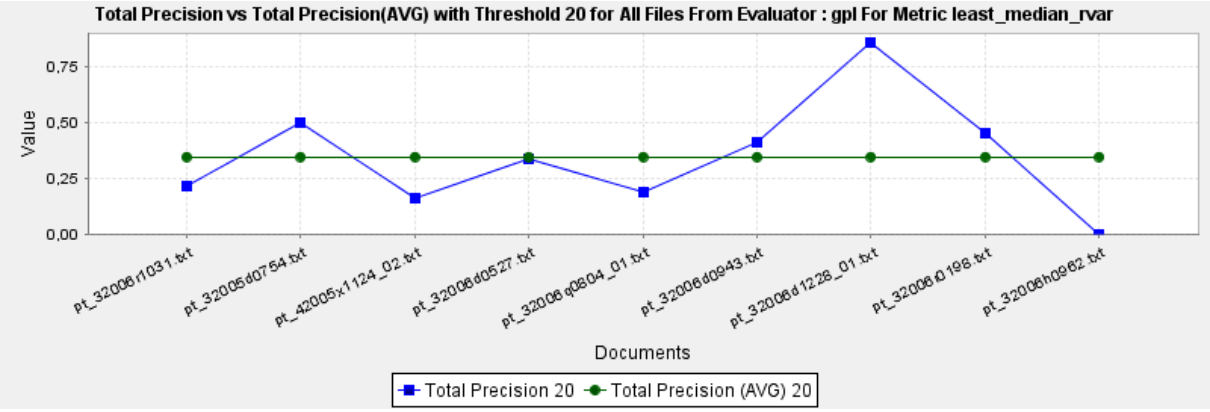


Figura 8.19 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median MI, com o limite 5

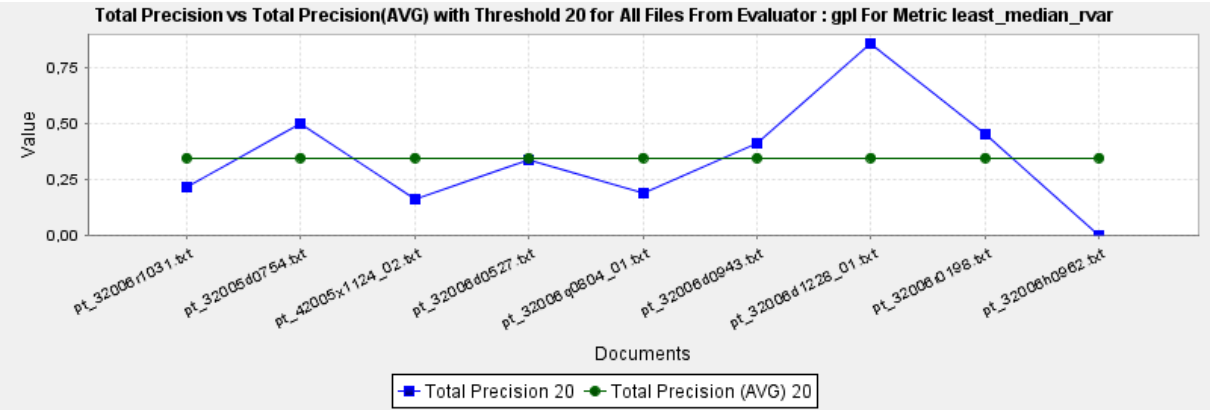


Figura 8.20 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median MI, com o limite 20

8.8 Tabela da Precisão Total Média para todas as Medidas resultante da Avaliação dos documentos em português pelo Avaliador Prof. Gabriel Lopes

Metric	Prec. Avg (5)	Prec. Avg (10)	Prec. Avg (15)	Prec. Avg (20)
least_bubbled_median_rvar	0,516666667	0,483289242	0,453106153	0,414740896
least_bubbled_phisquare	0,6	0,644973545	0,657305657	0,653613824
phisquare	0,727777778	0,725	0,68026048	0,621251386
least_median_tf_idf	0,683333333	0,632451499	0,660758377	0,638938724
bubbled_phisquare	0,666666667	0,650925926	0,645719096	0,61327884
least_median_rvar	0,462962963	0,355202822	0,347985348	0,345351328
least_bubbled_median_mi	0,516666667	0,47808642	0,444120694	0,432757547
least_bubbled_tf_idf	0,861111111	0,710582011	0,651890085	0,64977531
least_median_phisquare	0,611111111	0,63505291	0,58006993	0,593688097
bubbled_mi	N/A	N/A	N/A	N/A
least_phisquare	0,683333333	0,63968254	0,618270618	0,59459922
least_median_mi	0,424074074	0,353968254	0,351628002	0,334064942
bubbled_rvar	N/A	N/A	N/A	N/A
least_tf_idf	0,638888889	0,660978836	0,640761091	0,645621202
least_bubbled_median_phisquare	0,622222222	0,613580247	0,62049062	0,626377422
least_bubbled_median_tf_idf	0,833333333	0,696604938	0,678927554	0,684558493
rvar	N/A	N/A	N/A	N/A
least_rvar	N/A	N/A	0,347354497	0,315756898
tf_idf	0,694444444	0,702469136	0,709427609	0,659259259
least_bubbled_rvar	N/A	N/A	N/A	N/A
mi	N/A	N/A	N/A	N/A
least_bubbled_mi	N/A	N/A	N/A	N/A
least_mi	N/A	N/A	N/A	0,347322555
bubbled_tf_idf	0,824074074	0,687654321	0,682299182	0,662905709

Tabela 8.28 - Precisão total média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

8.9 Tabela da Cobertura Média para todas as Medidas resultante da Avaliação dos documentos em português pelo Avaliador Prof. Gabriel Lopes

Metric	Recall Avg (5)	Recall Avg (10)	Recall Avg (15)	Recall Avg (20)
least_bubbled_median_rvar	0,055350608	0,088076416	0,110701215	0,133545601
least_bubbled_phisquare	0,129548253	0,213912785	0,292546953	0,334466145
phisquare	0,162332188	0,303927597	0,399484185	0,484566035
least_median_tf_idf	0,147956117	0,233989511	0,348686783	0,439054225
bubbled_phisquare	0,130895055	0,209667864	0,253438907	0,300759944
least_median_rvar	0,057282204	0,079072186	0,102677377	0,143163089
least_bubbled_median_mi	0,055350608	0,073986239	0,119675011	0,150109794
least_bubbled_tf_idf	0,161000026	0,24721641	0,320974881	0,373235768
least_median_phisquare	0,138441103	0,218055663	0,285497774	0,365331921
bubbled_mi	0,030149319	0,067083882	0,087351779	0,096611038
least_phisquare	0,141134706	0,233704048	0,309680781	0,355486665
least_median_mi	0,061528327	0,078817157	0,104421022	0,1321988
bubbled_rvar	0,036209925	0,060906163	0,086004977	0,091055482
least_tf_idf	0,140275652	0,245604161	0,347772559	0,463789118
least_bubbled_median_phisquare	0,136911887	0,234905856	0,292186886	0,352236805
least_bubbled_median_tf_idf	0,156718506	0,21896329	0,313311719	0,398012951
rvar	0,039408579	0,049795052	0,061161918	0,065992835
least_rvar	0,03733714	0,047218562	0,073261602	0,086627141
tf_idf	0,143491608	0,286616807	0,393796016	0,475884592
least_bubbled_rvar	0,036209925	0,065737081	0,080449422	0,106053286
mi	0	0,005555556	0,021753339	0,029160746
least_bubbled_mi	0,030149319	0,063380179	0,087351779	0,111103792
least_mi	0,020048309	0,05204948	0,076745718	0,086627141
bubbled_tf_idf	0,144480181	0,231274381	0,273156975	0,294149509

Tabela 8.29 - Cobertura média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

8.10 Gráficos das Precisões para o Avaliador Prof. Joaquim Ferreira da Silva para o documento pt_32006R0198.html

As seguintes figuras apresentam os gráficos com as precisões, cobertura e F-Measure, considerados mais demonstrativos e foram obtidas da análise dos resultados do avaliador Prof. Joaquim Ferreira da Silva para o documento pt_32006R0198.html[64]. Os gráficos mostram os valores de precisão para 5, 10, 15 e 20.

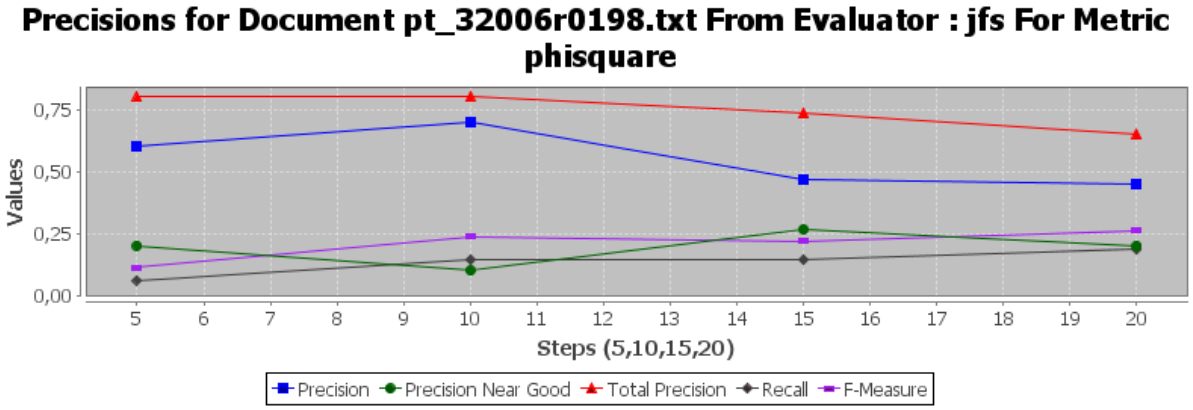


Figura 8.21 - Valores de Precisão, Cobertura e F-Measure para Phi-Square

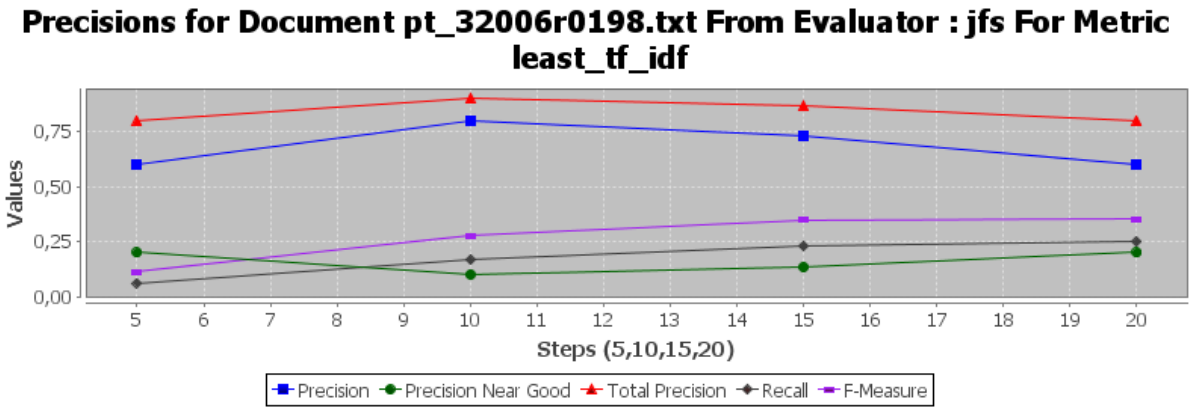


Figura 8.22 - Valores de Precisão, Cobertura e F-Measure para Least Tf-Idf

[64] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R0198:PT:NOT

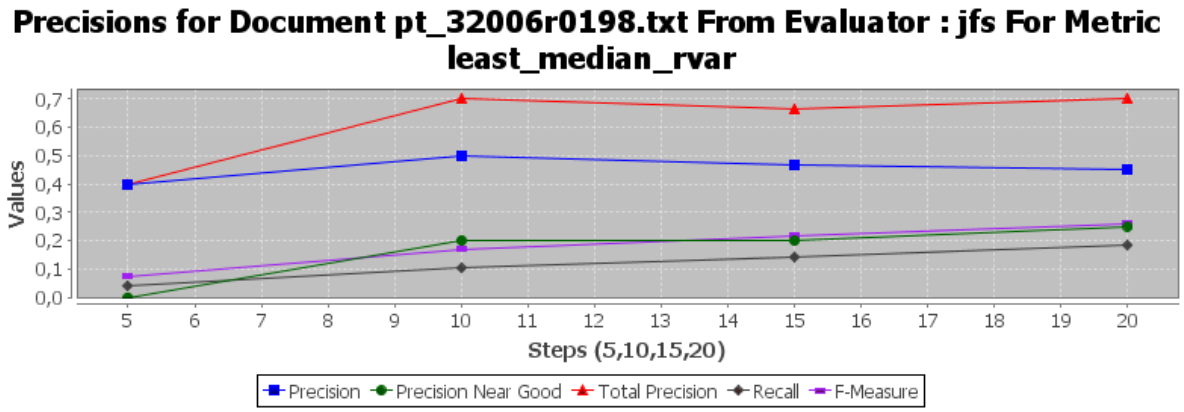


Figura 8.23 - Valores de Precisão, Cobertura e F-Measure para Least Median Rvar

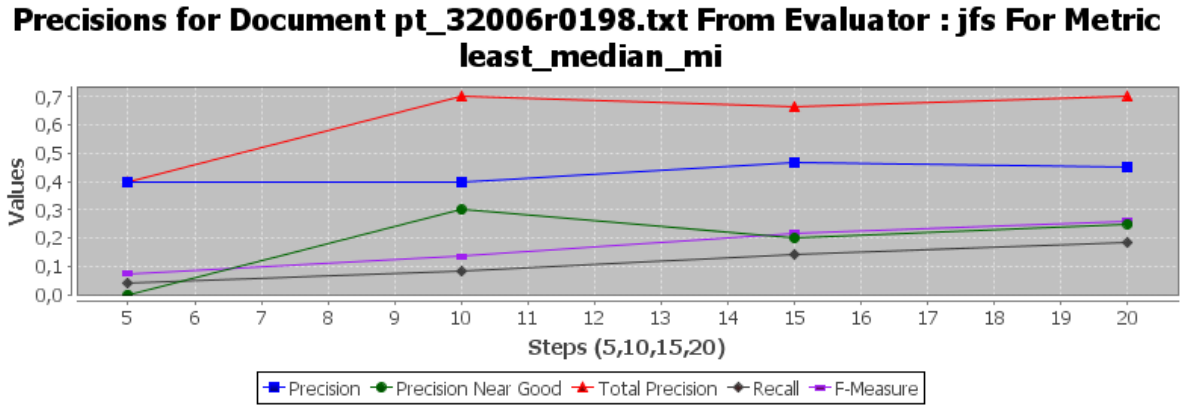


Figura 8.24 - Valores de Precisão, Cobertura e F-Measure para Least Median MI

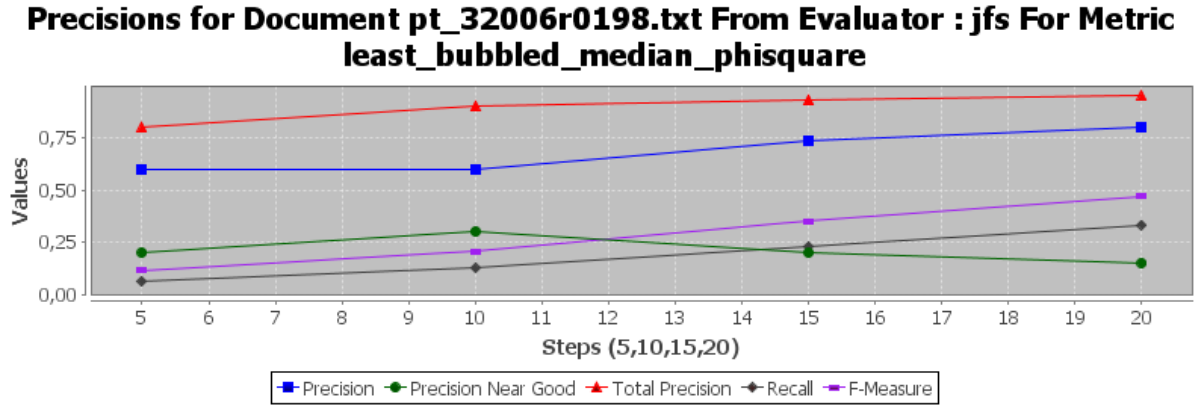


Figura 8.25 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Phi-Square

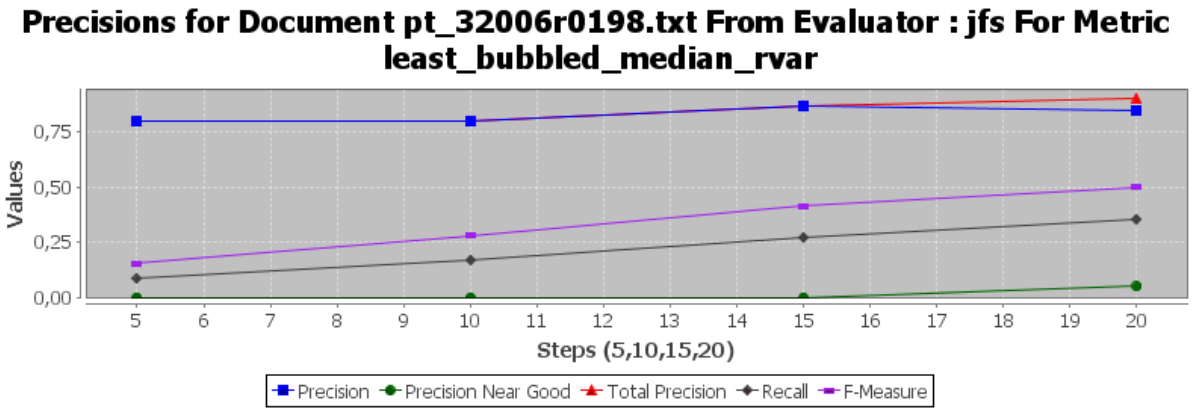


Figura 8.26 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Rvar

8.11 Gráficos da Precisão Total para todos os documentos em português avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva

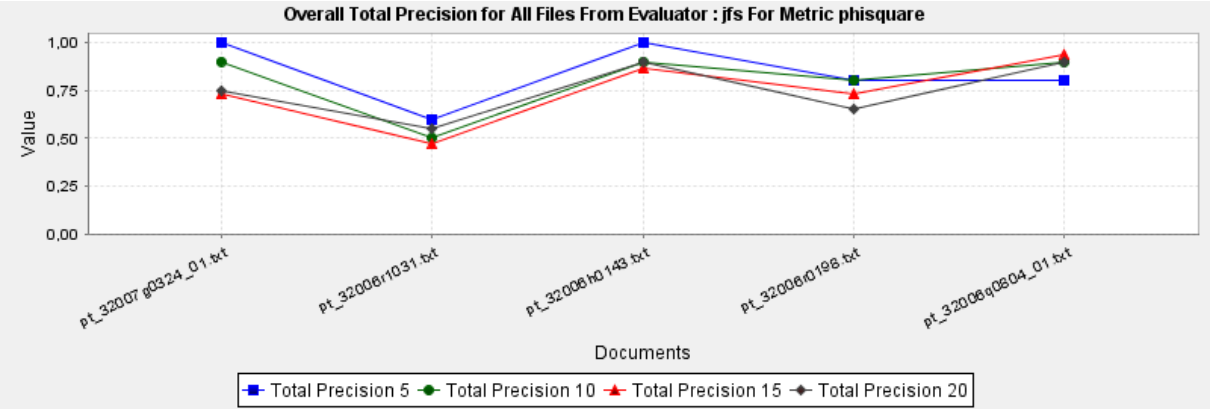


Figura 8.27 - Precisão total para todos os documentos em Português, para a medida Phi-Square

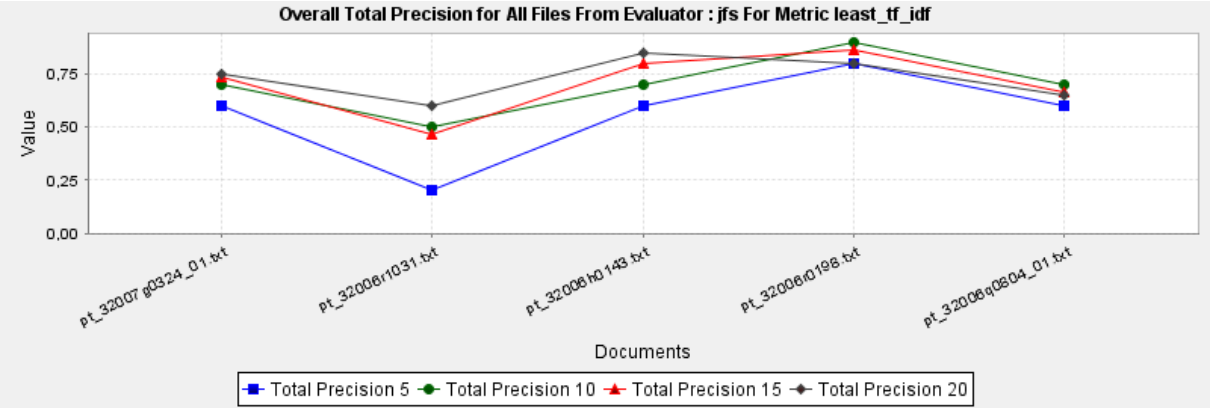


Figura 8.28 - Precisão total para todos os documentos em Português, para a medida Least Tf-Idf

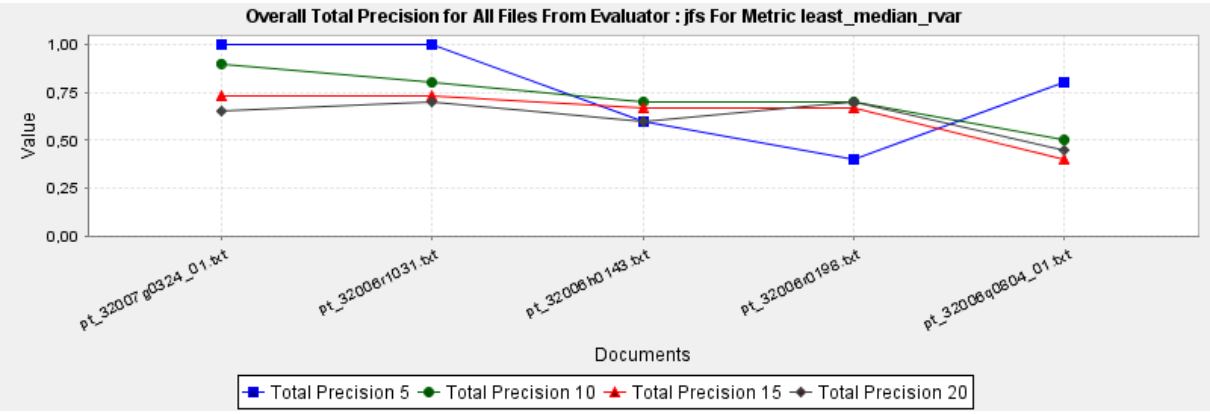


Figura 8.29 - Precisão total para todos os documentos em Português, para a medida Least Median Rvar

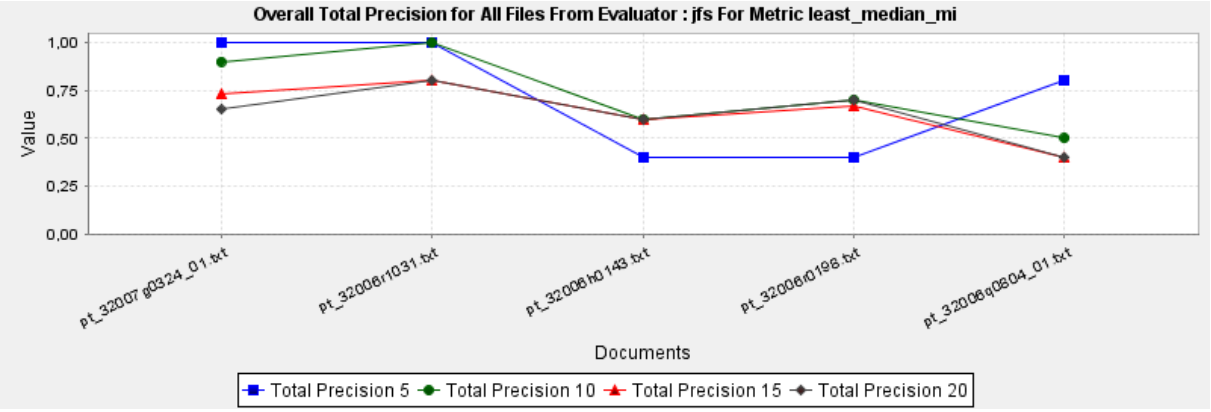


Figura 8.30 - Precisão total para todos os documentos em Português, para a medida Least Median MI

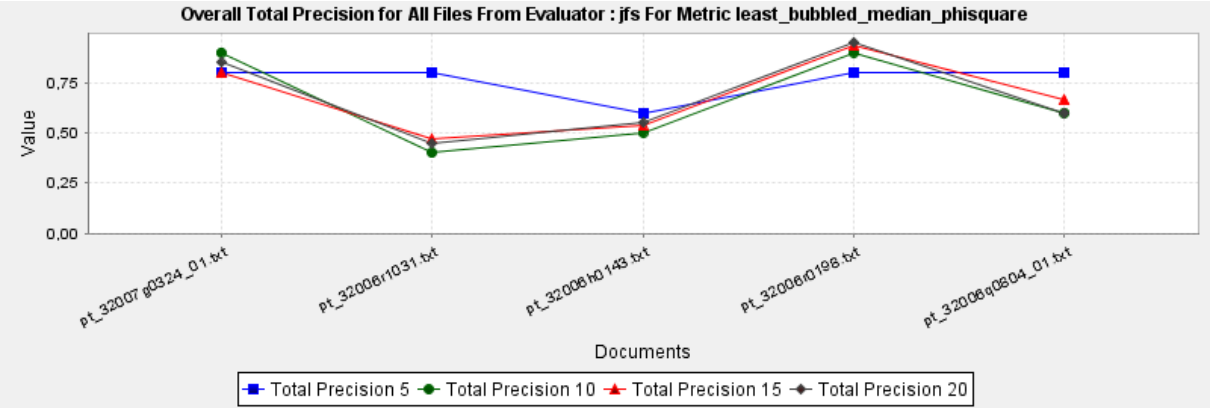


Figura 8.31 - Precisão total para todos os documentos em Português, para a medida Least Bubbled Median Phi-Square

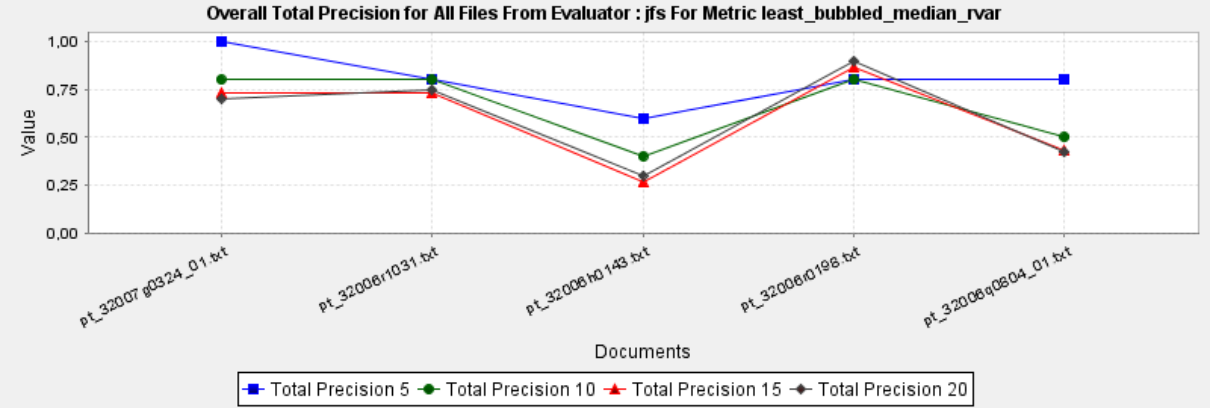


Figura 8.32 - Precisão total para todos os documentos em Português, para a medida Least Bubbled Median Rvar

8.12 Gráficos da Precisão Total versus Média da Precisão Total para todos os documentos em português avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva

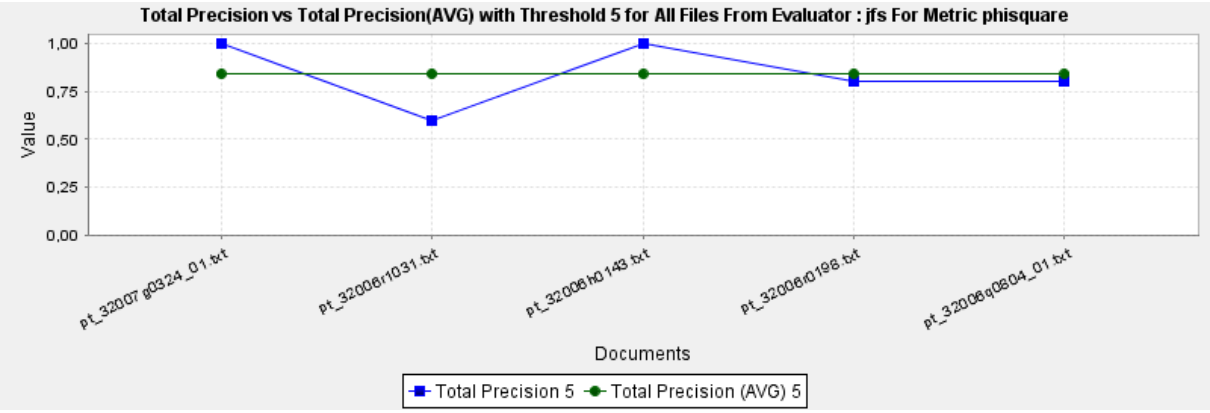


Figura 8.33 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 5

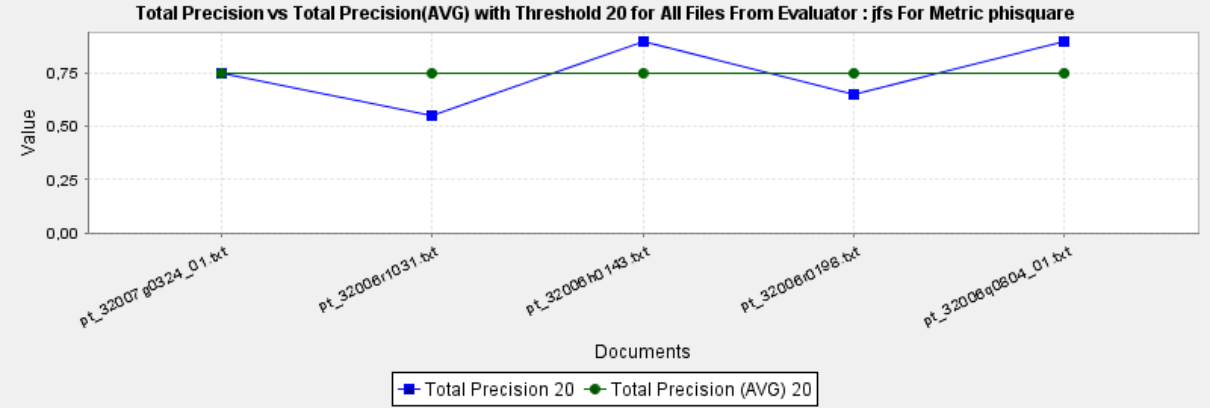


Figura 8.34 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 20

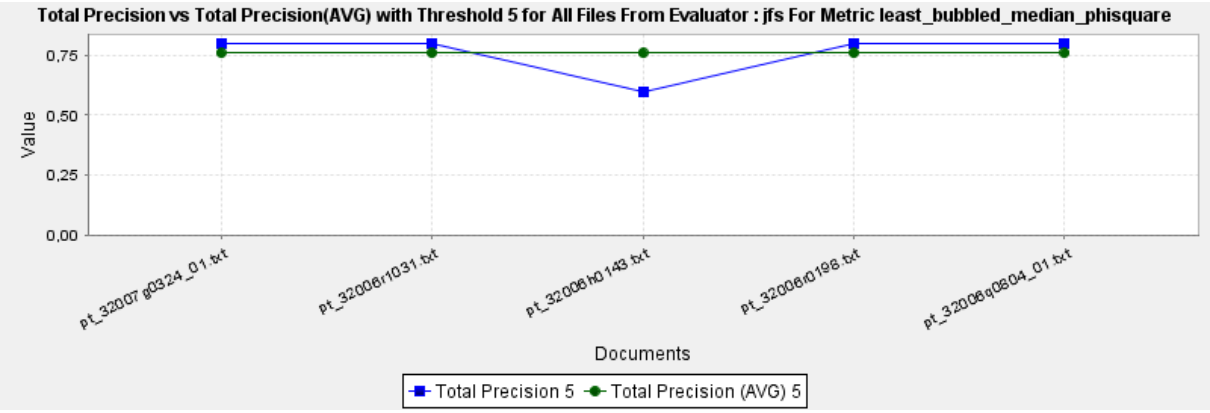


Figura 8.35 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 5

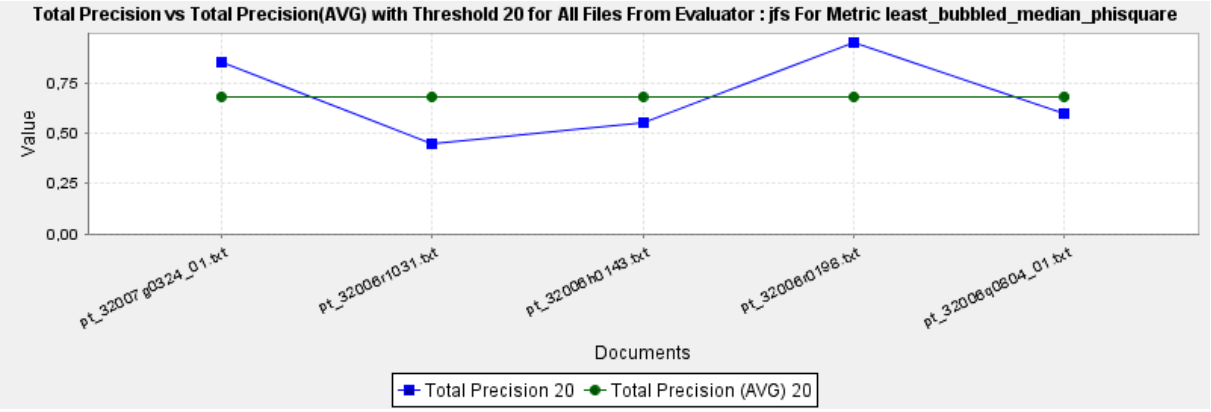


Figura 8.36 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 20

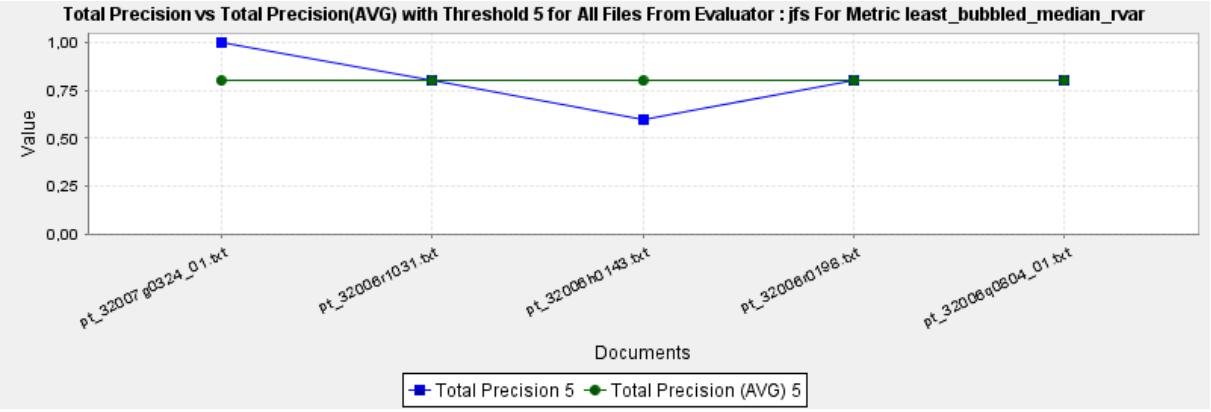


Figura 8.37 Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Rvar, com o limite **5**

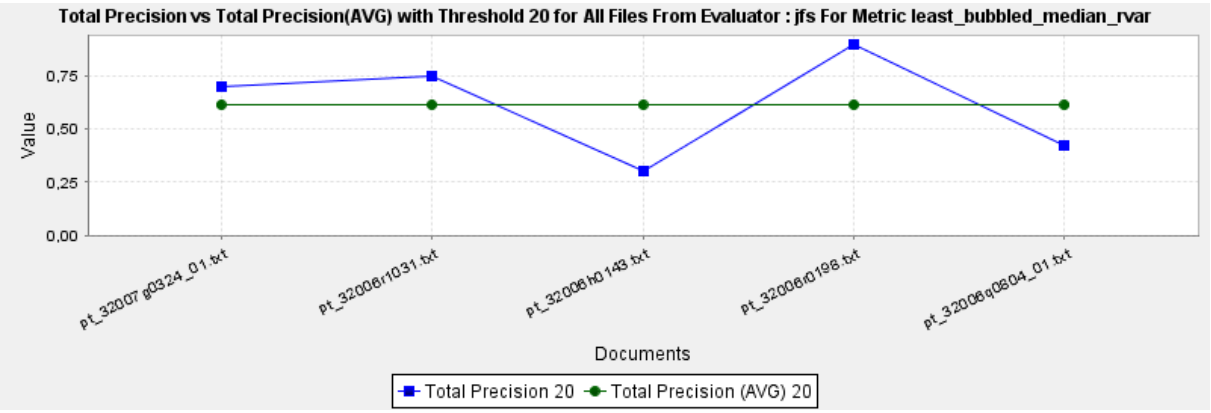


Figura 8.38 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Rvar, com o limite 20

8.13 Tabela da Precisão Total Média para todas as Medidas resultante da Avaliação dos documentos em português pelo Avaliador Prof. Joaquim Ferreira da Silva

Metric	Prec. Avg (5)	Prec. Avg (10)	Prec. Avg (15)	Prec. Avg (20)
least_bubbled_median_rvar	0,8	0,66	0,605714286	0,614210526
least_bubbled_phisquare	0,6	0,62	0,674285714	0,687041624
phisquare	0,84	0,8	0,746666667	0,75
least_median_tf_idf	0,8	0,757777778	0,77047619	0,754561404
bubbled_phisquare	0,68	0,713333333	0,73025641	0,71122291
least_median_rvar	0,76	0,72	0,64	0,62
least_bubbled_median_mi	0,8	0,733333333	0,687655678	0,640144479
least_bubbled_tf_idf	0,76	0,706746032	0,725128205	0,7
least_median_phisquare	0,8	0,78	0,693333333	0,722923977
bubbled_mi	0,733333333	0,688095238	0,593339993	0,610930736
least_phisquare	0,72	0,66	0,716666667	0,671176471
least_median_mi	0,72	0,74	0,64	0,63
bubbled_rvar	N/A	0,75	0,645244755	0,591486291
least_tf_idf	0,56	0,7	0,706666667	0,73
least_bubbled_median_phisquare	0,76	0,66	0,68	0,68
least_bubbled_median_tf_idf	0,84	0,755	0,731282051	0,739640769
rvar	N/A	N/A	N/A	N/A
least_rvar	0,533333333	0,545	0,523181818	0,572619048
tf_idf	0,84	0,8	0,773333333	0,76
least_bubbled_rvar	N/A	0,721428571	0,656153846	0,565445665
mi	N/A	N/A	N/A	N/A
least_bubbled_mi	0,693333333	0,681428571	0,598188478	0,618223443
least_mi	0,7	0,681904762	0,661038961	0,578073593
bubbled_tf_idf	0,84	0,745555556	0,758974359	0,689705882

Tabela 8.30 - Precisão total média, para todas as medidas, resultante da avaliação do Avaliador Prof. Joaquim Ferreira da Silva

8.14 Tabela da Cobertura Média para todas as Medidas resultante da Avaliação dos documentos em português pelo Avaliador Prof. Joaquim Ferreira da Silva

Metric	Recall Avg (5)	Recall Avg (10)	Recall Avg (15)	Recall Avg (20)
least_bubbled_median_rvar	0,089853115	0,146752468	0,197265355	0,26465666
least_bubbled_phisquare	0,067580933	0,123703382	0,174904714	0,222810826
phisquare	0,100914266	0,166227626	0,211752786	0,285856612
least_median_tf_idf	0,089009505	0,143892025	0,222477927	0,296408067
bubbled_phisquare	0,075517441	0,102841704	0,153136006	0,2102861
least_median_rvar	0,085279527	0,155097352	0,19441078	0,228846158
least_bubbled_median_mi	0,086235345	0,161533266	0,208879722	0,2561128
least_bubbled_tf_idf	0,078979109	0,124050755	0,183521494	0,248661167
least_median_phisquare	0,086121967	0,165327833	0,220316495	0,274752738
bubbled_mi	0,038609061	0,084980943	0,103635355	0,126764607
least_phisquare	0,082947363	0,144912433	0,201949148	0,255524195
least_median_mi	0,084681554	0,158861147	0,193804458	0,234336855
bubbled_rvar	0,038435374	0,084980943	0,106522893	0,126764607
least_tf_idf	0,062085921	0,135645273	0,208842305	0,291097308
least_bubbled_median_phisquare	0,080534448	0,137478892	0,198855961	0,255690645
least_bubbled_median_tf_idf	0,084028079	0,132274328	0,202742799	0,280074299
rvar	0,028571429	0,039795918	0,039795918	0,039795918
least_rvar	0,023129252	0,039289333	0,05765668	0,076311092
tf_idf	0,105169586	0,163619916	0,222977276	0,297081102
least_bubbled_rvar	0,038435374	0,084980943	0,106522893	0,127271192
mi	0	0,004081633	0,009637188	0,009637188
least_bubbled_mi	0,034527428	0,08089931	0,103635355	0,131526511
least_mi	0,020241714	0,043544652	0,054769142	0,079372316
bubbled_tf_idf	0,07977276	0,116000868	0,1750784	0,203763207

Tabela 8.31 - Cobertura média, para todas as medidas, resultante da avaliação do Avaliador Prof. Joaquim Ferreira da Silva

8.15 Cálculos da Estatística Kappa entre Prof. Joaquim Ferreira da Silva e o Prof. Gabriel Lopes para o documento en_32006Q804_01.html

8.15.1 Kappa para a Medida Phi-Square.

Este cálculo refere-se à medida *Phi-Square* para o documento en_32006Q804_01.html[65]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	11	0	0	0	0	11
	Near Good Descriptor	2	1	1	0	0	4
	Bad Descriptor	1	0	9	0	0	10
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	14	1	10	0	0	25

Tabela 8.32 - Matriz Confusão de Resultados Verificados para Phi-Square

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	6,2	0,4	4,4	0,0	0,0	11,0
	Near Good Descriptor	2,2	0,2	1,6	0,0	0,0	4,0
	Bad Descriptor	5,6	0,4	4,0	0,0	0,0	10,0
	Unkown	0,0	0,0	0,0	0,0	0,0	0,0
	No Evaluation	0,0	0,0	0,0	0,0	0,0	0,0
	Column Total	14,0	1,0	10,0	0,0	0,0	25,0

Tabela 8.33 - Matriz Confusão de Resultados Esperados para Phi-Square

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,727520435967302, o que dá aproximadamente 72.75% de concordância.

[65] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.15.2 Kappa para a Medida Least Tf-Idf

Este cálculo refere-se à medida *Least Tf-Idf* para o documento en_32006Q804_01.html[66]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	7	3	0	0	0	10
	Near Good Descriptor	4	1	1	0	0	6
	Bad Descriptor	1	0	8	0	0	9
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	12	4	9	0	0	25

Tabela 8.34 - Matriz Confusão de Resultados Verificados para Least Tf-Idf

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	4,8	1,6	3,6	0	0	10
	Near Good Descriptor	2,88	0,96	2,16	0	0	6
	Bad Descriptor	4,32	1,44	3,24	0	0	9
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	12	4	9	0	0	25

Tabela 8.35 - Matriz Confusão de Resultados Esperados para Least Tf-Idf

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,4375, o que dá aproximadamente 43.75% de concordância.

[66] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.15.3 Kappa para a Medida Least Median Rvar

Este cálculo refere-se à medida *Least Median Rvar* para o documento en_32006Q804_01.html[67]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	3	3	1	2	0	9
	Near Good Descriptor	0	1	4	3	0	8
	Bad Descriptor	0	0	8	0	0	8
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	4	13	5	0	25

Tabela 8.36 - Matriz Confusão de Resultados Verificados para Least Median Rvar

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	1,08	1,44	4,68	1,8	0	9
	Near Good Descriptor	0,96	1,28	4,16	1,6	0	8
	Bad Descriptor	0,96	1,28	4,16	1,6	0	8
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	4	13	5	0	25

Tabela 8.37 - Matriz Confusão de Resultados Esperados para Least Median Rvar

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,296536796536796, o que dá aproximadamente 26.65% de concordância.

[67] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.15.4 Kappa para a Medida Least Median MI

Este cálculo refere-se à medida *Least Median MI* para o documento en_32006Q804_01.html[68]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	3	3	2	2	0	10
	Near Good Descriptor	0	1	4	3	0	8
	Bad Descriptor	0	0	7	0	0	7
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	4	13	5	0	25

Tabela 8.38- Matriz Confusão de Resultados Verificados para Least Median MI

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	1,2	1,6	5,2	2	0	10
	Near Good Descriptor	0,96	1,28	4,16	1,6	0	8
	Bad Descriptor	0,84	1,12	3,64	1,4	0	7
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	4	13	5	0	25

Tabela 8.39 - Matriz Confusão de Resultados Esperados para Least Median MI

Com estas duas matrizes, o valor de Kappa, ver secção 2.8.3 sobre o cálculo da estatística, obtido é de 0,258474576271186, o que dá aproximadamente 25.84% de concordância.

[68] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.15.5 Kappa para a Medida Least Bubbled Median Phi-Square

Este cálculo refere-se à medida *Least Bubbled Median Phi-Square* para o documento en_32006Q804_01.html[69]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	8	2	1	0	0	11
	Near Good Descriptor	2	0	0	0	0	2
	Bad Descriptor	1	0	11	0	0	12
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	11	2	12	0	0	25

Tabela 8.40 - Matriz Confusão de Resultados Verificados para Least Bubbled Median Phi-Square

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	4,84	0,88	5,28	0	0	11
	Near Good Descriptor	0,88	0,16	0,96	0	0	2
	Bad Descriptor	5,28	0,96	5,76	0	0	12
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	11	2	12	0	0	25

Tabela 8.41 - Matriz Confusão de Resultados Esperados para Least Bubbled Median Phi-Square

Com estas duas matrizes, o valor de Kappa obtido, ver secção 2.8.3 sobre o cálculo da estatística, é de 0,578651685393258, o que dá aproximadamente 57.86% de concordância.

[69] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.15.6 Kappa para a Medida Least Bubbled Median Rvar

Este cálculo refere-se à medida *Least Bubbled Median Rvar* para o documento en_32006Q804_01.html[70]

Seja considerado o seguinte:

- Avaliador 1: Prof. Joaquim Ferreira da Silva.
- Avaliador 2: Prof. Gabriel Lopes.

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	3	1	0	1	0	5
	Near Good Descriptor	0	0	4	6	0	10
	Bad Descriptor	0	0	10	0	0	10
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	1	14	7	0	25

Tabela 8.42 - Matriz Confusão de Resultados Verificados para Least Bubbled Median Rvar

		Avaliador 2					
		Good Descriptor	**Near Good Descriptor**	**Bad Descriptor**	**Unkown**	**No Evaluation**	**Line Total**
Avaliador 1	**Good Descriptor**	0,6	0,2	2,8	1,4	0	5
	Near Good Descriptor	1,2	0,4	5,6	2,8	0	10
	Bad Descriptor	1,2	0,4	5,6	2,8	0	10
	Unkown	0	0	0	0	0	0
	No Evaluation	0	0	0	0	0	0
	Column Total	3	1	14	7	0	25

Tabela 8.43 - Matriz Confusão de Resultados Esperados para Least Bubbled Median Rvar

Com estas duas matrizes, o valor de Kappa obtido, ver secção 2.8.3 sobre o cálculo da estatística, é 0.34783 o que dá aproximadamente 34.78% de concordância.

[70] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006Q0804%2801%29:EN:HTML

8.16 Lista de Termos Avaliados pelo Avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html

8.16.1 Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
governing board	0,016368033116676	good topic descriptor
governing	0,014533005724990	bad descriptor
chairperson	0,010633486245839	good topic descriptor
bureau	0,006954830301350	good topic descriptor
director	0,004513219266702	good topic descriptor
founding regulation	0,004090793192082	good topic descriptor
founding	0,004090793192082	bad descriptor
centre	0,003606283277149	good topic descriptor
director of the centre	0,003272569769547	good topic descriptor
voting	0,002891409949613	bad descriptor
motion	0,002196500393209	good topic descriptor
if the chairperson	0,002045295373861	bad descriptor
meeting	0,001901388889910	good topic descriptor
attend	0,001811246676773	bad descriptor
members	0,001787372645332	near good descriptor
minutes	0,001772238243083	good topic descriptor
he / she	0,001687973498046	bad descriptor
members of the governing	0,001636220104502	bad descriptor
members of the governing board	0,001636220104502	good topic descriptor
unable to attend	0,001636220104502	bad descriptor
majority	0,001636220104502	bad descriptor
vice-chairpersons	0,001636220104502	good topic descriptor
meetings of the governing board	0,001636220104502	good topic descriptor
meetings of the governing	0,001636220104502	bad descriptor
development of vocational training	0,001293200838982	good topic descriptor

Tabela 8.44 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Phi-Square

8.16.2 Least Tf-Idf

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
chairperson	0,029851088353419	good topic descriptor
governing	0,029590879977958	bad descriptor
bureau	0,023731661781725	good topic descriptor
bureau and the governing	0,023731661781725	bad descriptor
governing board and the bureau	0,023731661781725	good topic descriptor
founding	0,013959801048074	bad descriptor
director	0,013267150379297	good topic descriptor
director and deputy director	0,013267150379297	good topic descriptor
chairperson or the director	0,013267150379297	good topic descriptor
centre	0,009292295675709	good topic descriptor
director of the centre	0,009292295675709	good topic descriptor
voting	0,008844766919532	bad descriptor
members of the governing	0,007828313677225	bad descriptor
members	0,007828313677225	near good descriptor
chairperson considers that a motion	0,007739171054590	bad descriptor
motion may impede the governing	0,007739171054590	bad descriptor
motion	0,007739171054590	good topic descriptor
minutes	0,005706481375529	good topic descriptor
attend	0,005614391842917	bad descriptor
majority of members	0,005583920419229	near good descriptor
chairperson and the vice-chairpersons	0,005583920419229	good topic descriptor
majority	0,005583920419229	bad descriptor
vice-chairpersons and members	0,005583920419229	near good descriptor
majority of its members	0,005583920419229	near good descriptor
vice-chairpersons	0,005583920419229	good topic descriptor

Tabela 8.45 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Least Tf-Idf

8.16.3 Least Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	17,000000000000000	good topic descriptor
simultaneously	14,000000000000000	bad descriptor
admissibility	**13,000000000000000**	**unkonwn**
countersigned	13,000000000000000	bad descriptor
far-reaching	12,000000000000000	bad descriptor
appointments	12,000000000000000	near good descriptor
ascertained	11,000000000000000	bad descriptor
explanation	**11,000000000000000**	**unkonwn**
nominations	**11,000000000000000**	**near good descriptor**
nominations and appointments	11,000000000000000	near good descriptor
secretariat	11,000000000000000	near good descriptor
scrutineers	**11,000000000000000**	**unkonwn**
medium-term	11,000000000000000	bad descriptor
vice-chairs	11,000000000000000	good topic descriptor
precedence	**10,000000000000000**	**unkonwn**
indication	10,000000000000000	bad descriptor
chairperson	9,488692799006760	good topic descriptor
chairperson and countersigned	9,488692799006760	bad descriptor
substance	9,000000000000000	bad descriptor
convening	9,000000000000000	bad descriptor
seniority	**9,000000000000000**	**unkonwn**
forthwith	9,000000000000000	bad descriptor
postponed	9,000000000000000	bad descriptor
therefrom	9,000000000000000	bad descriptor
deletion therefrom	8,500000000000000	bad descriptor

Tabela 8.46 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Least Median Rvar

8.16.4 Least Median MI

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	63,673145221654230	good topic descriptor
simultaneously	52,436707829597600	bad descriptor
admissibility	**48,691228698912056**	**unkonwn**
countersigned	48,691228698912056	bad descriptor
far-reaching	44,945749568226520	bad descriptor
appointments	44,945749568226520	near good descriptor
ascertained	41,200270437540970	bad descriptor
explanation	**41,200270437540970**	**unkonwn**
nominations	41,200270437540970	near good descriptor
nominations and appointments	41,200270437540970	near good descriptor
secretariat	41,200270437540970	near good descriptor
scrutineers	**41,200270437540970**	**unkonwn**
medium-term	41,200270437540970	bad descriptor
vice-chairs	41,200270437540970	good topic descriptor
chairperson	40,800226351661344	good topic descriptor
chairperson and countersigned	40,800226351661344	bad descriptor
precedence	**37,454791306855430**	**unkonwn**
indication	37,454791306855430	bad descriptor
correspondence	37,056135788244060	bad descriptor
substance	33,709312176169890	bad descriptor
convening	33,709312176169890	bad descriptor
seniority	**33,709312176169890**	**unkonwn**
forthwith	33,709312176169890	bad descriptor
postponed	33,709312176169890	bad descriptor
therefrom	33,709312176169890	bad descriptor

Tabela 8.47 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Least Median MI

8.16.5 Least Bubbled Median Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
chairperson	0,116968348704232	good topic descriptor
governments	0,075066368633285	bad descriptor
governing	0,061417937972688	bad descriptor
bureau	0,041728981808101	good topic descriptor
vice-chairpersons	0,041724438596906	good topic descriptor
governing board and the bureau	0,034121076651493	good topic descriptor
founding	0,032726345536657	bad descriptor
bureau and the governing	0,030708968986344	bad descriptor
vice-chairs	0,026998166150939	good topic descriptor
motions	0,023633032442703	good topic descriptor
meetings	0,023119033314776	good topic descriptor
chairperson considers that a motion	0,020256884950889	bad descriptor
motion may impede the governing	0,020256884950889	bad descriptor
motion	0,020256884950889	good topic descriptor
meeting	0,020229154150429	good topic descriptor
governing the centre between meetings	0,020229154150429	good topic descriptor
motions that the governing	0,018568811204981	bad descriptor
attendance	0,017722382430834	near good descriptor
voting	0,017348459697676	bad descriptor
chairperson and the vice-chairpersons	0,017180651186961	good topic descriptor
centre between meetings	0,016113607939238	bad descriptor
meetings of the governing	0,015894335403908	bad descriptor
vice-chairs of the governing	0,014726272445967	bad descriptor
chairperson shall close the meeting	0,014449395821735	bad descriptor
attendance at meetings	0,014177905944667	near good descriptor

Tabela 8.48 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Least Bubbled Median Phi-Square

8.16.6 Least Bubbled Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	16,999999999999996	good topic descriptor
simultaneously	14,000000000000000	bad descriptor
admissibility	**13,000000000000000**	**unkonwn**
countersigned	13,000000000000000	bad descriptor
far-reaching	12,000000000000000	bad descriptor
ascertained	11,000000000000000	bad descriptor
explanation	**11,000000000000000**	**unkonwn**
vice-chairs	10,999999999999998	good topic descriptor
chairperson	9,488692799006760	good topic descriptor
chairperson and countersigned	9,488692799006760	bad descriptor
seniority	**9,000000000000000**	**unkonwn**
forthwith	9,000000000000000	bad descriptor
postponed	9,000000000000000	bad descriptor
precedence	**8,655720030369995**	**unkonwn**
deletion	**8,000000000000000**	**unkonwn**
absolute majority	8,000000000000000	near good descriptor
absolute	8,000000000000000	bad descriptor
majority	8,000000000000000	bad descriptor
founding	7,999999999999998	bad descriptor
chairperson thinks	7,332171708323406	bad descriptor
revised	7,000000000000000	bad descriptor
besides	7,000000000000000	bad descriptor
speaker	**7,000000000000000**	**unkonwn**
validly	7,000000000000000	bad descriptor
figures	**7,000000000000000**	**unkonwn**

Tabela 8.49 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento en_32006Q804_01.html na medida Least Bubbled Median Rvar

8.17 Lista de Termos Avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html

8.17.1 Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
governing board	0,016368033116676	good topic descriptor
governing	0,014533005724990	bad descriptor
chairperson	0,010633486245839	good topic descriptor
bureau	0,006954830301350	good topic descriptor
director	0,004513219266702	good topic descriptor
founding regulation	0,004090793192082	good topic descriptor
founding	0,004090793192082	bad descriptor
centre	0,003606283277149	near good descriptor
director of the centre	0,003272569769547	good topic descriptor
voting	0,002891409949613	bad descriptor
motion	0,002196500393209	near good descriptor
if the chairperson	0,002045295373861	bad descriptor
meeting	0,001901388889910	good topic descriptor
attend	0,001811246676773	bad descriptor
members	0,001787372645332	near good descriptor
minutes	0,001772238243083	bad descriptor
he / she	0,001687973498046	bad descriptor
members of the governing	0,001636220104502	bad descriptor
members of the governing board	0,001636220104502	good topic descriptor
unable to attend	0,001636220104502	bad descriptor
majority	0,001636220104502	near good descriptor
vice-chairpersons	0,001636220104502	good topic descriptor
meetings of the governing board	0,001636220104502	good topic descriptor
meetings of the governing	0,001636220104502	bad descriptor
development of vocational training	0,001293200838982	good topic descriptor

Tabela 8.50 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Phi-Square

8.17.2 Least Tf-Idf

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
chairperson	0,029851088353419	good topic descriptor
governing	0,029590879977958	bad descriptor
bureau	0,023731661781725	good topic descriptor
bureau and the governing	0,023731661781725	bad descriptor
governing board and the bureau	0,023731661781725	near good descriptor
founding	0,013959801048074	bad descriptor
director	0,013267150379297	good topic descriptor
director and deputy director	0,013267150379297	good topic descriptor
chairperson or the director	0,013267150379297	near good descriptor
centre	0,009292295675709	near good descriptor
director of the centre	0,009292295675709	good topic descriptor
voting	0,008844766919532	bad descriptor
members of the governing	0,007828313677225	bad descriptor
members	0,007828313677225	near good descriptor
chairperson considers that a motion	0,007739171054590	bad descriptor
motion may impede the governing	0,007739171054590	bad descriptor
motion	0,007739171054590	near good descriptor
minutes	0,005706481375529	bad descriptor
attend	0,005614391842917	bad descriptor
majority of members	0,005583920419229	good topic descriptor
chairperson and the vice-chairpersons	0,005583920419229	good topic descriptor
majority	0,005583920419229	near good descriptor
vice-chairpersons and members	0,005583920419229	good topic descriptor
majority of its members	0,005583920419229	good topic descriptor
vice-chairpersons	0,005583920419229	good topic descriptor

Tabela 8.51 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Least Tf-Idf

8.17.3 Least Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	17,000000000000000	good topic descriptor
simultaneously	14,000000000000000	bad descriptor
admissibility	13,000000000000000	near good descriptor
countersigned	13,000000000000000	bad descriptor
far-reaching	12,000000000000000	bad descriptor
appointments	12,000000000000000	good topic descriptor
ascertained	11,000000000000000	bad descriptor
explanation	11,000000000000000	near good descriptor
nominations	11,000000000000000	good topic descriptor
nominations and appointments	11,000000000000000	good topic descriptor
secretariat	11,000000000000000	near good descriptor
scrutineers	11,000000000000000	good topic descriptor
medium-term	11,000000000000000	good topic descriptor
vice-chairs	11,000000000000000	good topic descriptor
precedence	10,000000000000000	near good descriptor
indication	10,000000000000000	near good descriptor
chairperson	9,488692799006760	good topic descriptor
chairperson and countersigned	9,488692799006760	bad descriptor
substance	9,000000000000000	near good descriptor
convening	9,000000000000000	bad descriptor
seniority	9,000000000000000	good topic descriptor
forthwith	9,000000000000000	near good descriptor
postponed	9,000000000000000	near good descriptor
therefrom	9,000000000000000	bad descriptor
deletion therefrom	8,500000000000000	bad descriptor

Tabela 8.52 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Least Median Rvar

8.17.4 Least Median MI

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	63,673145221654230	good topic descriptor
simultaneously	52,436707829597600	bad descriptor
admissibility	48,691228698912056	near good descriptor
countersigned	48,691228698912056	bad descriptor
far-reaching	44,945749568226520	bad descriptor
appointments	44,945749568226520	good topic descriptor
ascertained	41,200270437540970	bad descriptor
explanation	41,200270437540970	near good descriptor
nominations	41,200270437540970	good topic descriptor
nominations and appointments	41,200270437540970	good topic descriptor
secretariat	41,200270437540970	near good descriptor
scrutineers	41,200270437540970	good topic descriptor
medium-term	41,200270437540970	good topic descriptor
vice-chairs	41,200270437540970	good topic descriptor
chairperson	40,800226351661344	good topic descriptor
chairperson and countersigned	40,800226351661344	bad descriptor
precedence	37,454791306855430	near good descriptor
indication	37,454791306855430	near good descriptor
correspondence	37,056135788244060	good topic descriptor
substance	33,709312176169890	near good descriptor
convening	33,709312176169890	bad descriptor
seniority	33,709312176169890	good topic descriptor
forthwith	33,709312176169890	near good descriptor
postponed	33,709312176169890	near good descriptor
therefrom	33,709312176169890	bad descriptor

Tabela 8.53 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Least Median MI

8.17.5 Least Bubbled Median Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
chairperson	0,116968348704232	good topic descriptor
governments	0,075066368633285	good topic descriptor
governing	0,061417937972688	bad descriptor
bureau	0,041728981808101	good topic descriptor
vice-chairpersons	0,041724438596906	good topic descriptor
governing board and the bureau	0,034121076651493	near good descriptor
founding	0,032726345536657	bad descriptor
bureau and the governing	0,030708968986344	bad descriptor
vice-chairs	0,026998166150939	good topic descriptor
motions	0,023633032442703	good topic descriptor
meetings	0,023119033314776	good topic descriptor
chairperson considers that a motion	0,020256884950889	bad descriptor
motion may impede the governing	0,020256884950889	bad descriptor
motion	0,020256884950889	near good descriptor
meeting	0,020229154150429	good topic descriptor
governing the centre between meetings	0,020229154150429	bad descriptor
motions that the governing	0,018568811204981	bad descriptor
attendance	0,017722382430834	good topic descriptor
voting	0,017348459697676	bad descriptor
chairperson and the vice-chairpersons	0,017180651186961	good topic descriptor
centre between meetings	0,016113607939238	bad descriptor
meetings of the governing	0,015894335403908	bad descriptor
vice-chairs of the governing	0,014726272445967	bad descriptor
chairperson shall close the meeting	0,014449395821735	bad descriptor
attendance at meetings	0,014177905944667	good topic descriptor

Tabela 8.54 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Least Bubbled Median Phi-Square

8.17.6 Least Bubbled Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
vice-chairpersons	16,999999999999996	good topic descriptor
simultaneously	14,000000000000000	bad descriptor
admissibility	13,000000000000000	near good descriptor
countersigned	13,000000000000000	bad descriptor
far-reaching	12,000000000000000	bad descriptor
ascertained	11,000000000000000	bad descriptor
explanation	11,000000000000000	near good descriptor
vice-chairs	10,999999999999998	good topic descriptor
chairperson	9,488692799006760	good topic descriptor
chairperson and countersigned	9,488692799006760	bad descriptor
seniority	9,000000000000000	good topic descriptor
forthwith	9,000000000000000	near good descriptor
postponed	9,000000000000000	near good descriptor
precedence	8,655720030369995	near good descriptor
deletion	8,000000000000000	near good descriptor
absolute majority	8,000000000000000	good topic descriptor
absolute	8,000000000000000	bad descriptor
majority	8,000000000000000	near good descriptor
founding	7,999999999999998	bad descriptor
chairperson thinks	7,332171708323406	bad descriptor
revised	7,000000000000000	bad descriptor
besides	7,000000000000000	bad descriptor
speaker	7,000000000000000	near good descriptor
validly	7,000000000000000	near good descriptor
figures	7,000000000000000	near good descriptor

Tabela 8.55- Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html na medida Least Bubbled Median Rvar

8.18 Lista de Termos Apresentados aos Avaliadores para outras medidas

8.18.1 Rvar

Termo	Valor da Medida
she shall assist	1,00
one of the four categories	1,00
meeting and convene another	1,00
each belongs	1,00
admissibility of a motion	1,00
remarks are still	1,00
minutes of meetings	1,00
request to the notice	1,00
procedures to finalise	1,00
immediately bring any such request	1,00
a member may not	1,00
motion is put	1,00
chairperson thinks	1,00
his / her own initiative	1,00
which shall be made up	1,00
centre may be ascertained	1,00
divided into its several parts	1,00
board by written procedure	1,00
only for the meeting	1,00
if the chairperson considers	1,00
chairperson shall direct the proceedings	1,00
his own category	1,00
chairperson shall close	1,00
brief explanation	1,00
furthest from	1,00

Tabela 8.56 - Lista de Termos para a medida Rvar para o ficheiro en_32006Q804_01.html

8.18.2 MI

Termo	Valor da Medida
she shall assist	3,7454791
one of the four categories	3,7454791
meeting and convene another	3,7454791
each belongs	3,7454791
admissibility of a motion	3,7454791
remarks are still	3,7454791
minutes of meetings	3,7454791
request to the notice	3,7454791
procedures to finalise	3,7454791
immediately bring any such request	3,7454791
a member may not	3,7454791
motion is put	3,7454791
chairperson thinks	3,7454791
his / her own initiative	3,7454791
which shall be made up	3,7454791
centre may be ascertained	3,7454791
divided into its several parts	3,7454791
board by written procedure	3,7454791
only for the meeting	3,7454791
if the chairperson considers	3,7454791
chairperson shall direct the proceedings	3,7454791
his own category	3,7454791
chairperson shall close	3,7454791
brief explanation	3,7454791
furthest from	3,7454791

Tabela 8.57 - Lista de Termos para a medida MI para o ficheiro en_32006Q804_01.html

8.18.3 Tf-Idf

Termo	Valor da Medida
governing board	0,0558392
chairperson	0,0298511
governing	0,0295909
bureau	0,0237317
founding regulation	0,0139598
founding	0,0139598
director	0,0132672
director of the centre	0,0111678
centre	0,0092923
voting	0,0088448
members	0,0078283
motion	0,0077392
if the chairperson	0,0069799
minutes	0,0057065
attend	0,0056144
members of the governing	0,0055839
members of the governing board	0,0055839
unable to attend	0,0055839
majority	0,0055839
vice-chairpersons	0,0055839
meetings of the governing board	0,0055839
meetings of the governing	0,0055839
he / she	0,005528
his / her	0,0044224
development of vocational training	0,0044224

Tabela 8.58 - Lista de Termos para a medida Tf-Idf para o ficheiro en_32006Q804_01.html

8.19 Gráficos das Precisões para o Prof. Gabriel Lopes para o documento en_32006Q804_01.html

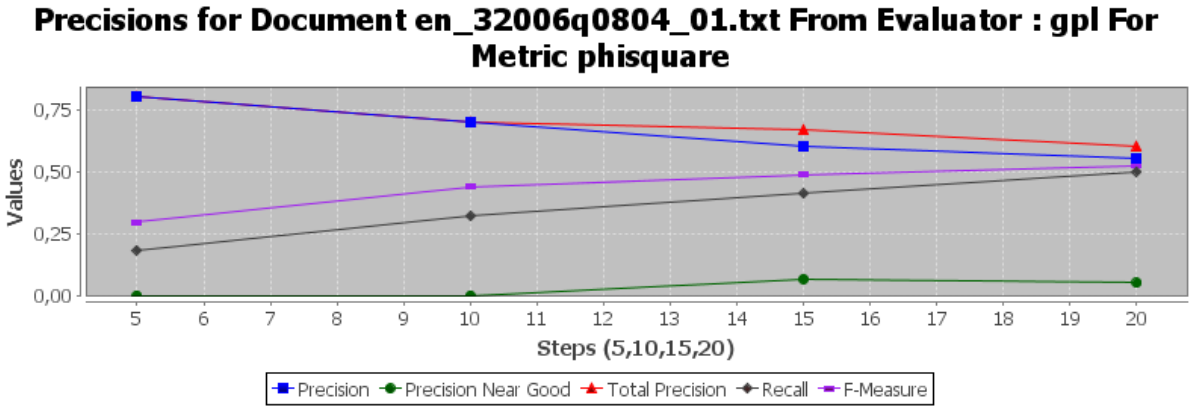


Figura 8.39 - Valores de Precisão, Cobertura e F-Measure para Phi-Square

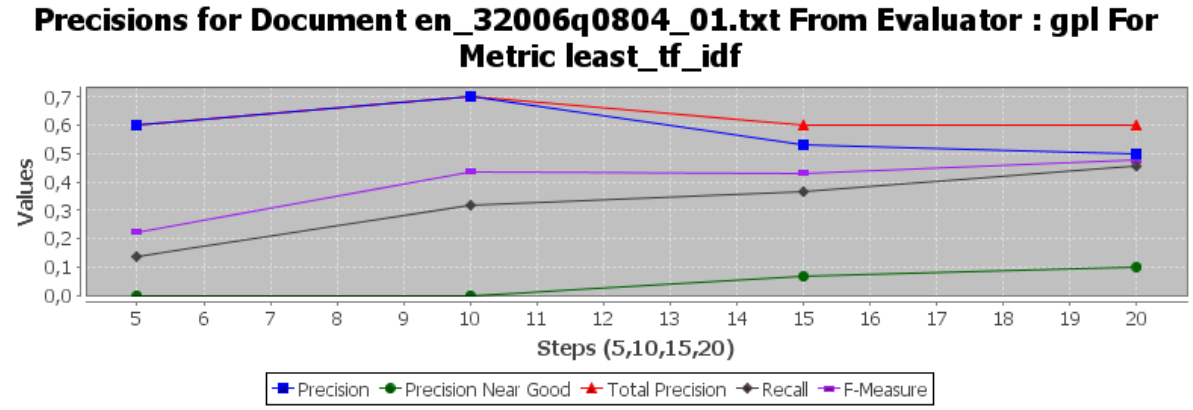


Figura 8.40 - Valores de Precisão, Cobertura e F-Measure para Least Tf-Idf

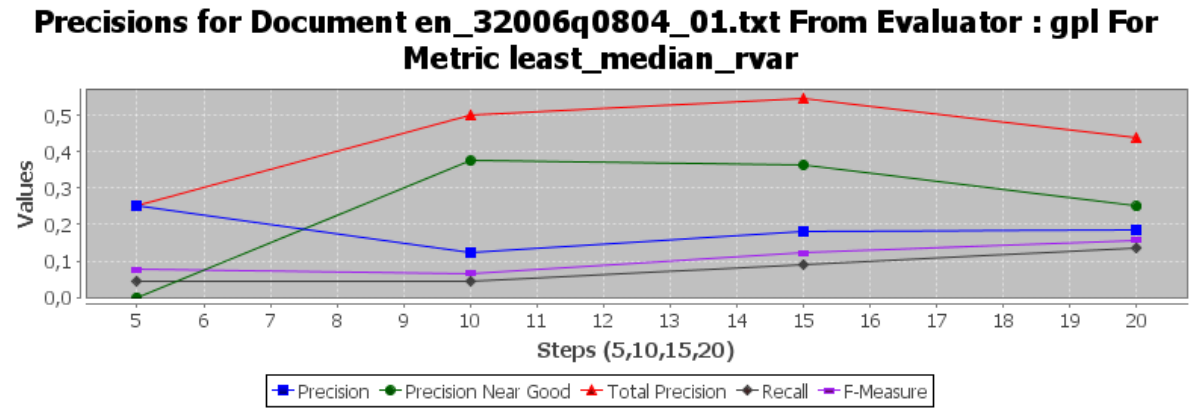


Figura 8.41 - Valores de Precisão, Cobertura e F-Measure para Least Median Rvar

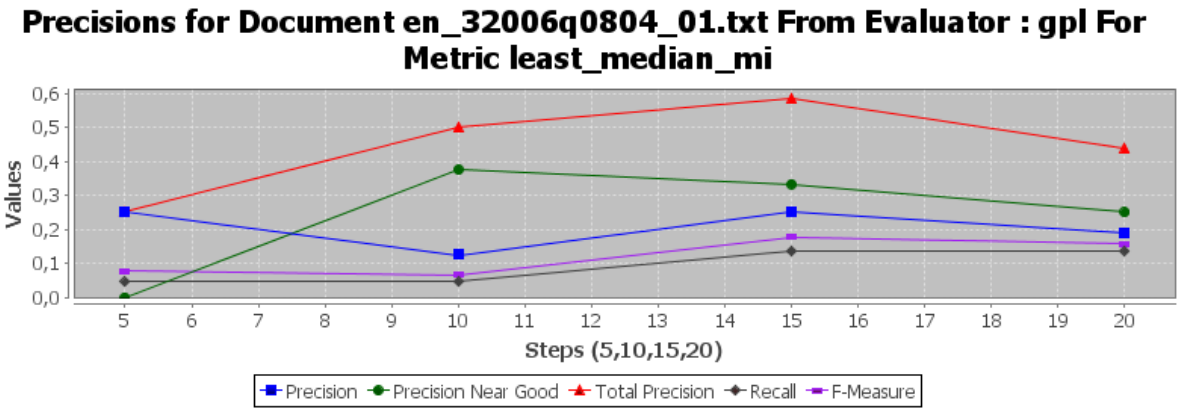


Figura 8.42 - Valores de Precisão, Cobertura e F-Measure para Least Median MI

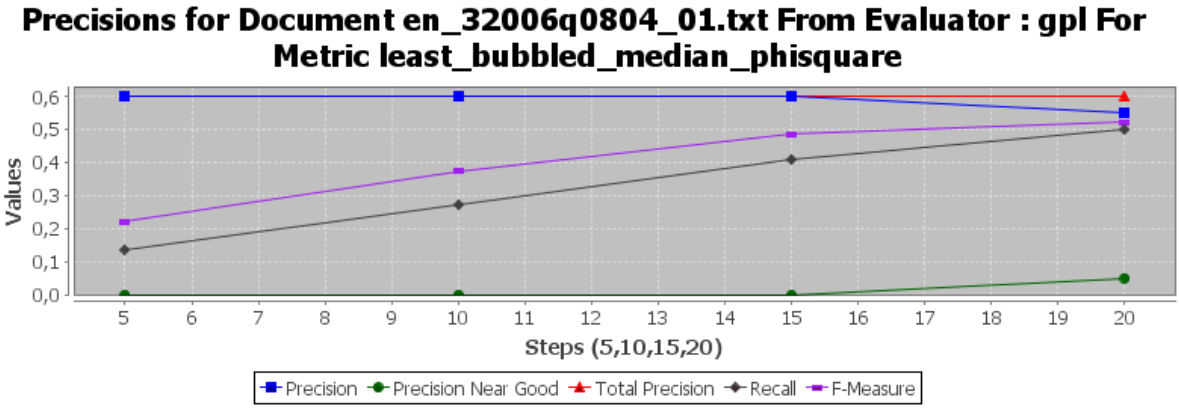


Figura 8.43 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Phi-Square

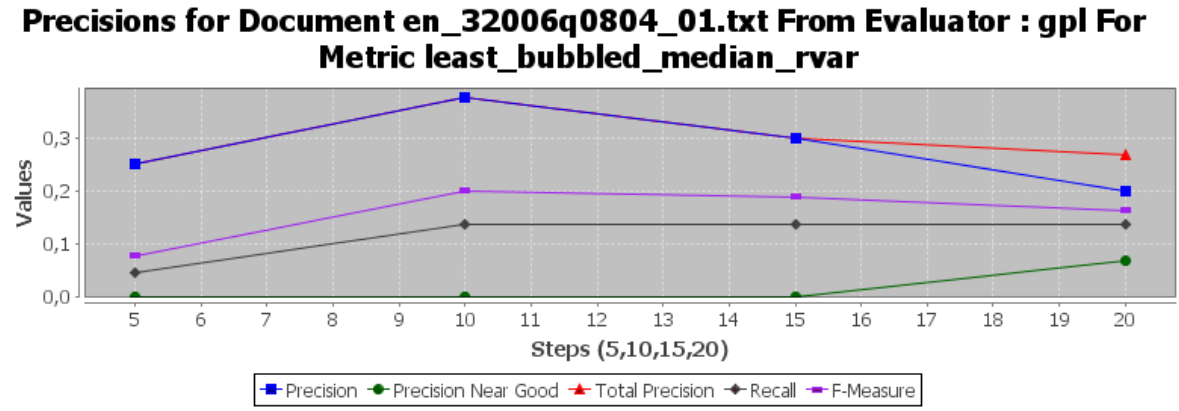


Figura 8.44 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Rvar

8.20 Gráficos da Precisão Total para todos os documentos em Inglês avaliados pelo Avaliador Prof. Gabriel Lopes

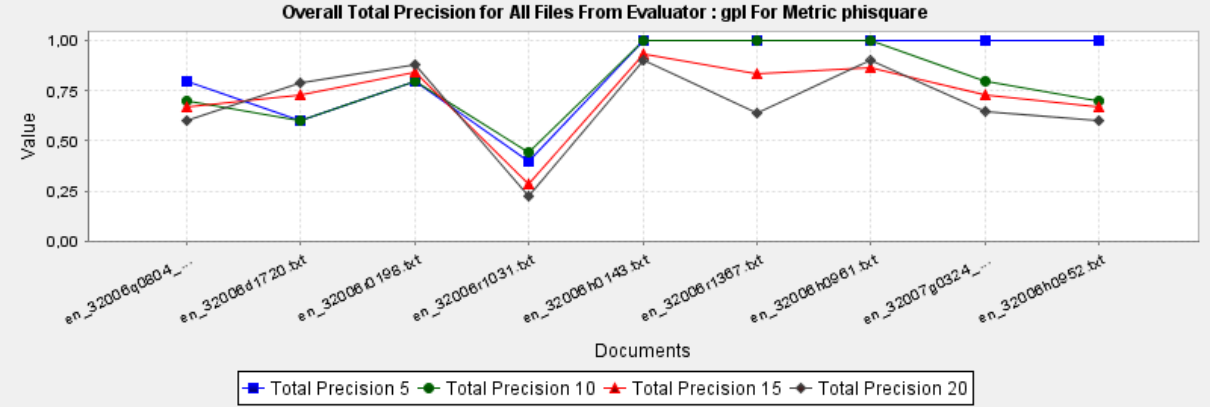


Figura 8.45 - Precisão total para todos os documentos em Inglês, para a medida Phi-Square

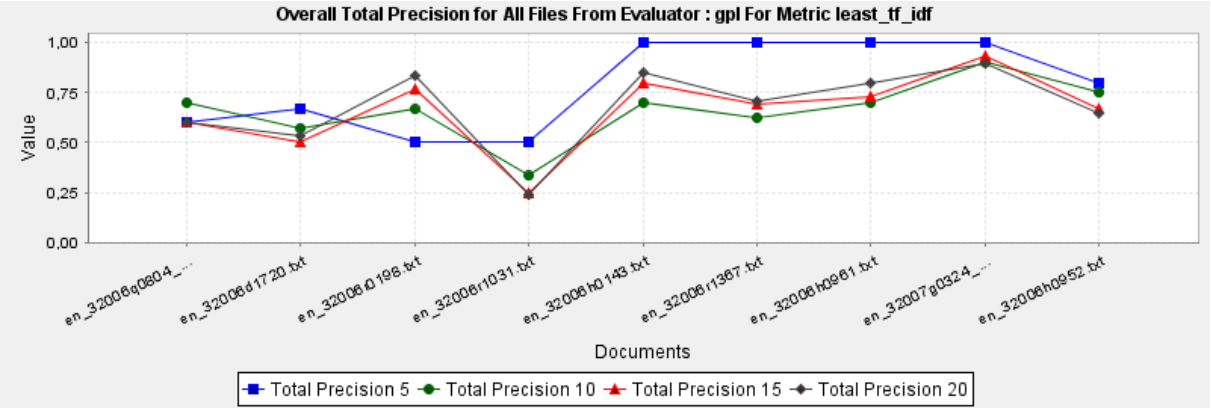


Figura 8.46 - Precisão total para todos os documentos em Inglês, para a medida Least Tf-Idf

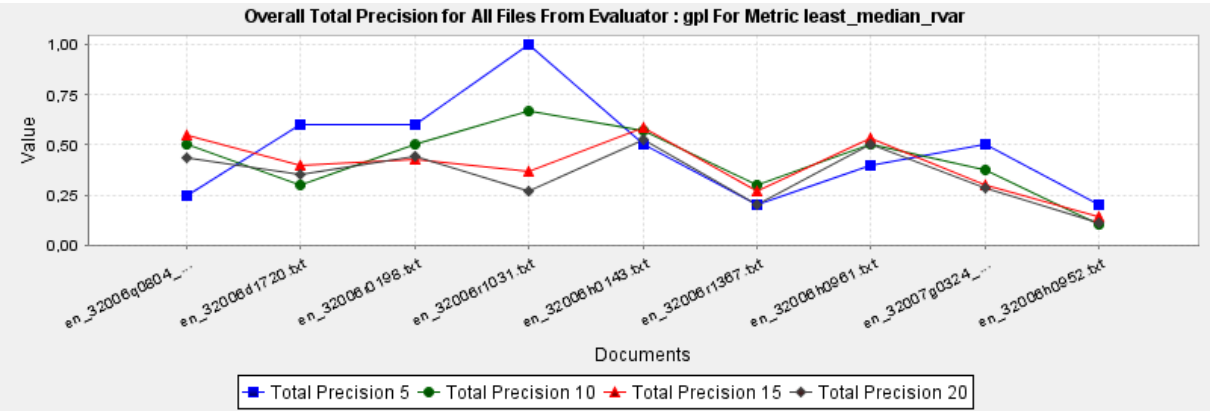


Figura 8.47- Precisão total para todos os documentos em Inglês, para a medida Least Median Rvar

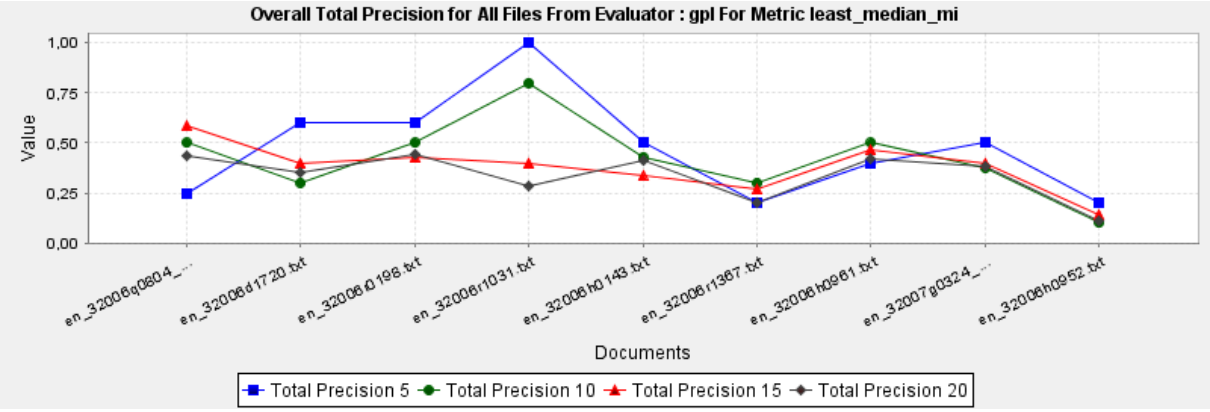


Figura 8.48 - Precisão total para todos os documentos em Inglês, para a medida Least Median MI

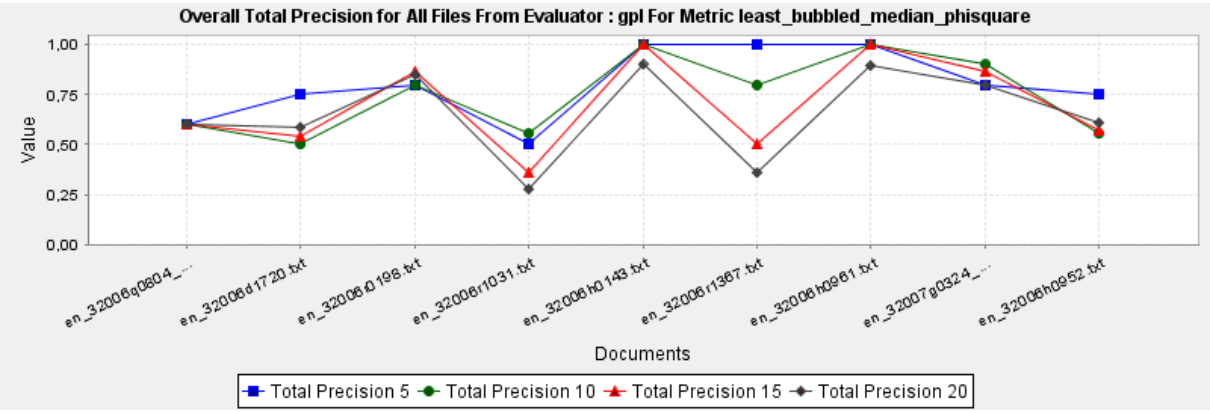


Figura 8.49 - Precisão total para todos os documentos em Inglês, para a medida Least Bubbled Median Phi-Square.

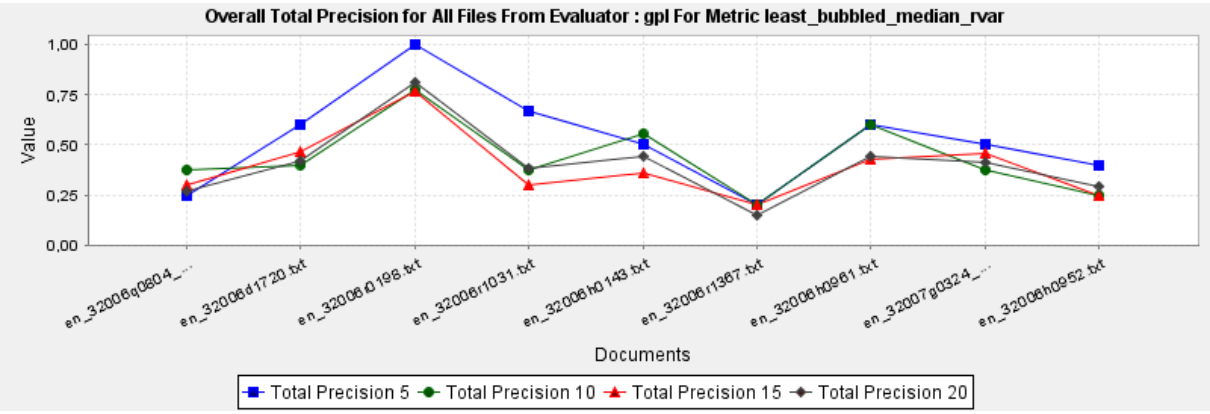


Figura 8.50 - Precisão total para todos os documentos em Inglês, para a medida Least Bubbled Median Rvar

8.21 Gráficos da Precisão Total versus Média da Precisão Total para todos os documentos em inglês avaliados pelo Avaliador Prof. Gabriel Lopes

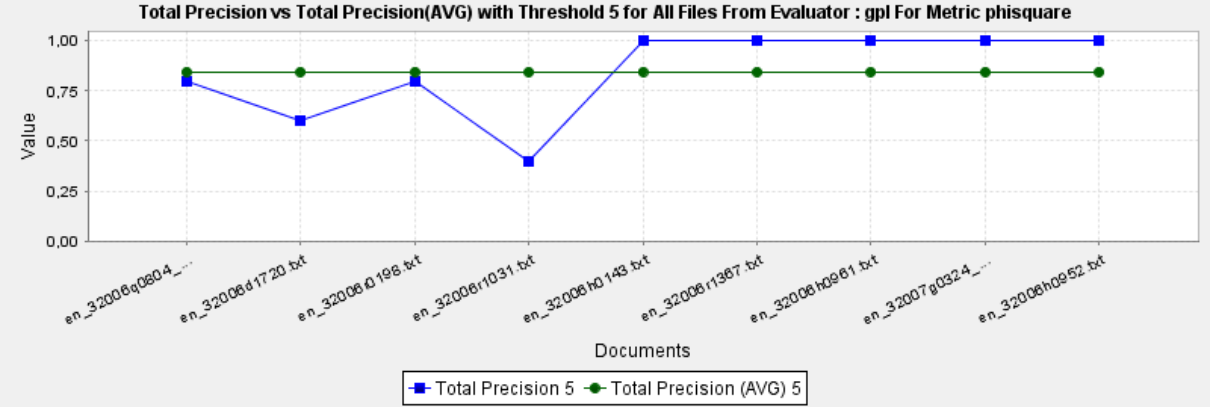


Figura 8.51 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 5

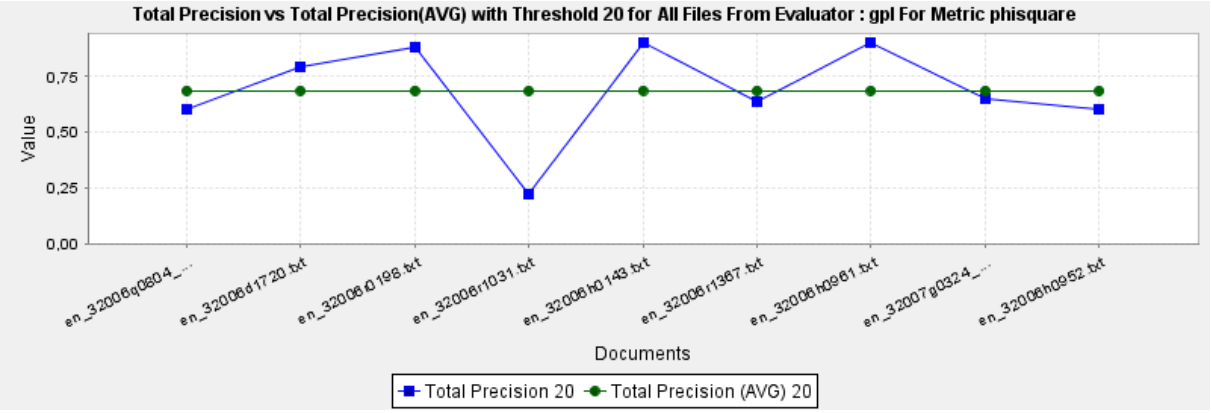


Figura 8.52 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 20

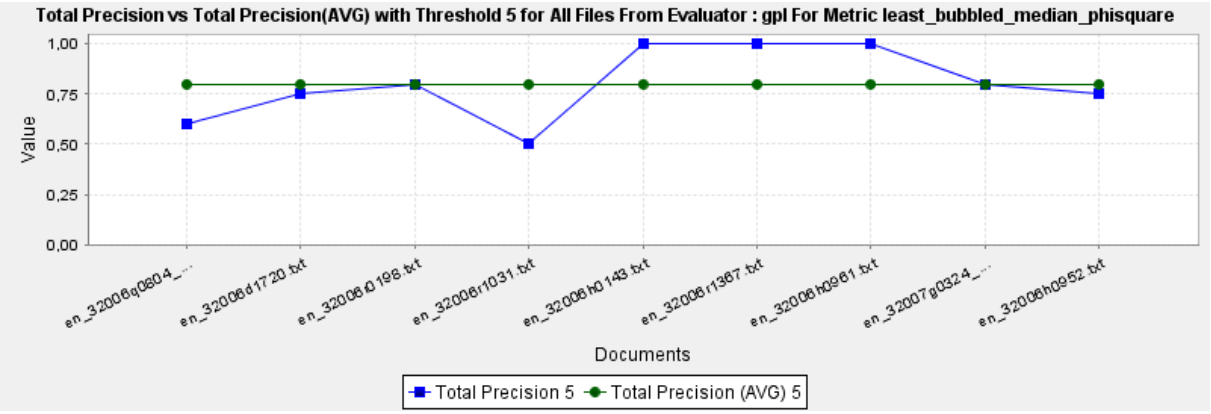


Figura 8.53 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 5

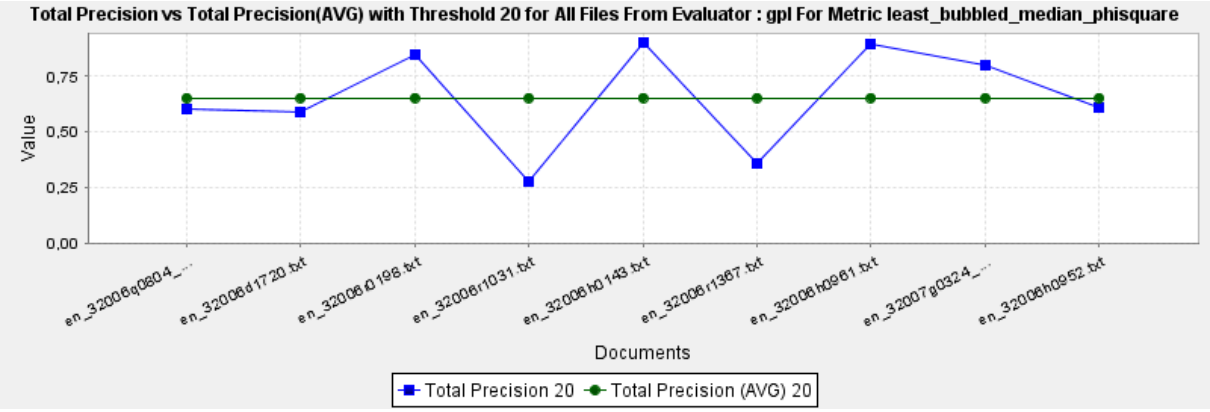


Figura 8.54 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 20

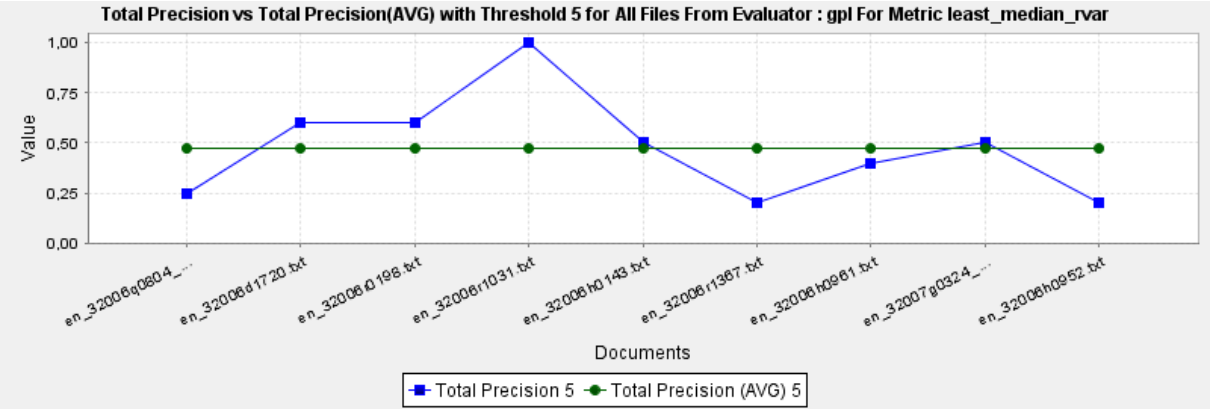


Figura 8.55 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 5

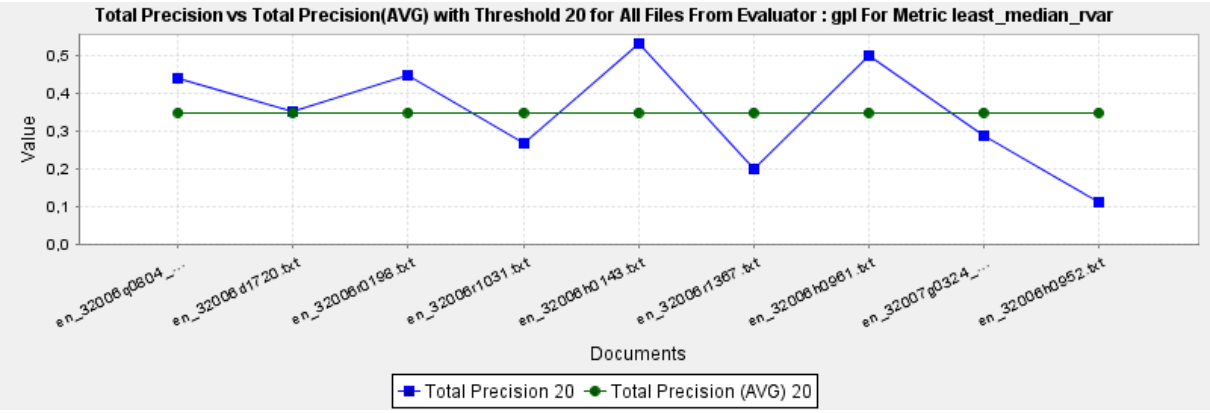


Figura 8.56 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 20

8.22 Tabela da Precisão Total Média para todas as Medidas resultante da Avaliação dos documentos em Inglês pelo Avaliador Prof. Gabriel Lopes

Metric	Precision Avg (5)	Precision Avg (10)	Precision Avg (15)	Precision Avg (20)
least_bubbled_median_rvar	0,524074074	0,434259259	0,392572243	0,403289547
least_bubbled_phisquare	0,82962963	0,742813051	0,678510379	0,620382866
phisquare	0,844444444	0,782716049	0,729466829	0,686712498
least_median_tf_idf	0,805555556	0,783201058	0,653825804	0,664581388
bubbled_phisquare	0,777777778	0,679012346	0,609279609	0,605546451
least_median_rvar	0,472222222	0,423677249	0,395983646	0,347205364
least_bubbled_median_mi	0,461111111	0,486816578	0,431826507	0,432492172
least_bubbled_tf_idf	0,846296296	0,65617284	0,652247752	0,646540077
least_median_phisquare	0,872222222	0,777777778	0,737932438	0,703376906
bubbled_mi	N/A	0,340388007	0,31957672	0,29761396
least_phisquare	0,82962963	0,759259259	0,754704555	0,690600685
least_median_mi	0,472222222	0,422619048	0,38015873	0,338466951
bubbled_rvar	N/A	0,282848325	0,304761905	0,313486976
least_tf_idf	0,785185185	0,660714286	0,660541311	0,677737645
least_bubbled_median_phisquare	0,8	0,745679012	0,7000407	0,653222654
least_bubbled_median_tf_idf	0,822222222	0,694973545	0,638071188	0,620031702
rvar	N/A	N/A	N/A	N/A
least_rvar	0,277777778	0,363492063	0,318903319	0,282814408
tf_idf	0,844444444	0,744444444	0,705575906	0,670454459
least_bubbled_rvar	N/A	0,282848325	0,307407407	0,35172217
mi	N/A	N/A	N/A	N/A
least_bubbled_mi	N/A	N/A	0,357936508	0,372949735
least_mi	0,17962963	0,282010582	0,303787879	0,28545991
bubbled_tf_idf	0,816666667	0,685185185	0,624809141	0,636980376

Tabela 8.59 - Precisão total média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

8.23 Tabela da Cobertura Média para todas as Medidas resultante da Avaliação dos documentos em Inglês pelo Avaliador Prof. Gabriel Lopes

Metric	Recall Avg (5)	Recall Avg (10)	Recall Avg (15)	Recall Avg (20)
least_bubbled_median_rvar	0,052743101	0,099584688	0,118747989	0,158558013
least_bubbled_phisquare	0,13788786	0,234784885	0,317835461	0,368746225
phisquare	0,141469168	0,289430085	0,356307435	0,447504494
least_median_tf_idf	0,128102638	0,283137812	0,345048678	0,428216983
bubbled_phisquare	0,102161457	0,16600486	0,203763398	0,248751844
least_median_rvar	0,033474497	0,066413619	0,096205324	0,115926344
least_bubbled_median_mi	0,042998306	0,102290596	0,128210248	0,155573516
least_bubbled_tf_idf	0,137812835	0,233742107	0,300462717	0,37629192
least_median_phisquare	0,169735751	0,283020083	0,371241415	0,42505056
bubbled_mi	0,02844552	0,050135504	0,06932115	0,084758128
least_phisquare	0,172008832	0,27766294	0,36619091	0,40872939
least_median_mi	0,033474497	0,056751783	0,08371186	0,109508125
bubbled_rvar	0,026858218	0,053117707	0,077858908	0,077858908
least_tf_idf	0,134872012	0,252416243	0,362758999	0,483546939
least_bubbled_median_phisquare	0,118659513	0,241410312	0,340923484	0,407674418
least_bubbled_median_tf_idf	0,098238167	0,186441368	0,292841996	0,353721587
rvar	0,023458369	0,023458369	0,029013924	0,032982178
least_rvar	0,011303511	0,048548203	0,048548203	0,058072012
tf_idf	0,152579903	0,269577089	0,349380912	0,434776981
least_bubbled_rvar	0,026858218	0,053117707	0,083414464	0,098631855
mi	0,023458369	0,023458369	0,029013924	0,029013924
least_bubbled_mi	0,02844552	0,046768501	0,074876705	0,105531075
least_mi	0,00952381	0,04616725	0,050135504	0,057470762
bubbled_tf_idf	0,119597725	0,176129921	0,219410814	0,259063292

Tabela 8.60 Cobertura média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

8.24 Gráficos das Precisões para o Avaliador Prof. Joaquim Ferreira da Silva para o documento en_32006Q804_01.html

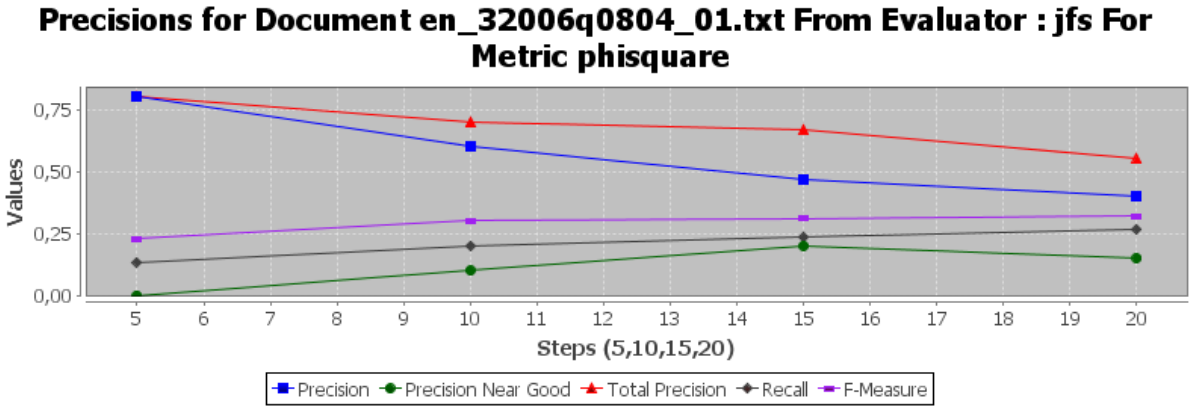


Figura 8.57 - Valores de Precisão, Cobertura e F-Measure para Phi-Square

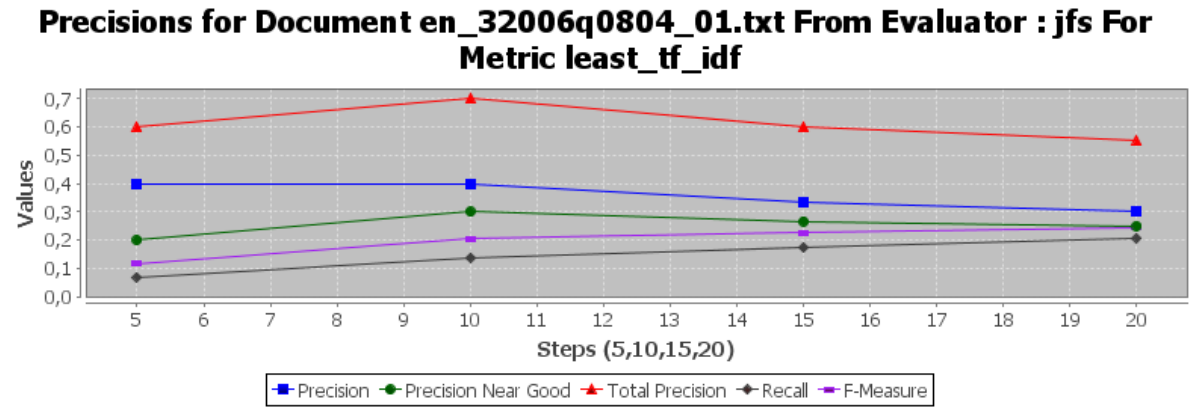


Figura 8.58 - Valores de Precisão, Cobertura e F-Measure para Least Tf-Idf

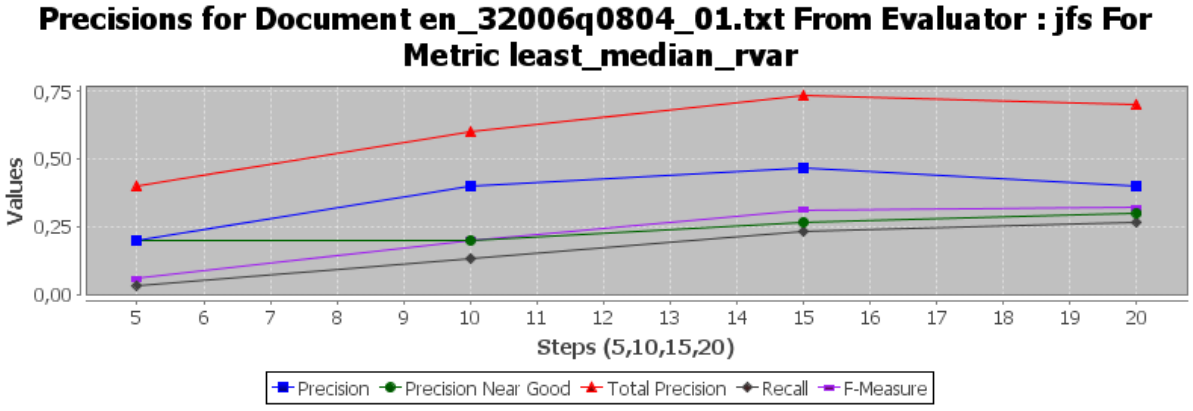


Figura 8.59 - Valores de Precisão, Cobertura e F-Measure para Least Median Rvar

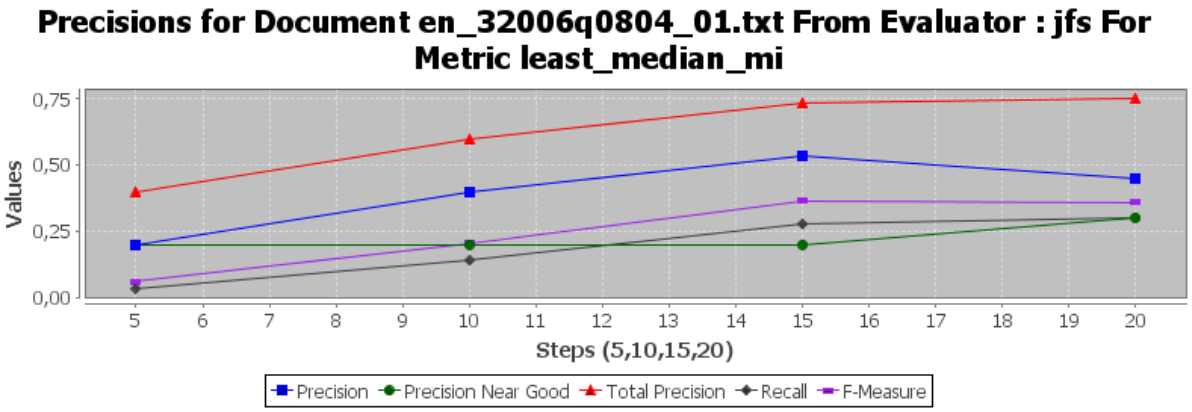


Figura 8.60 - Valores de Precisão, Cobertura e F-Measure para Least Median MI

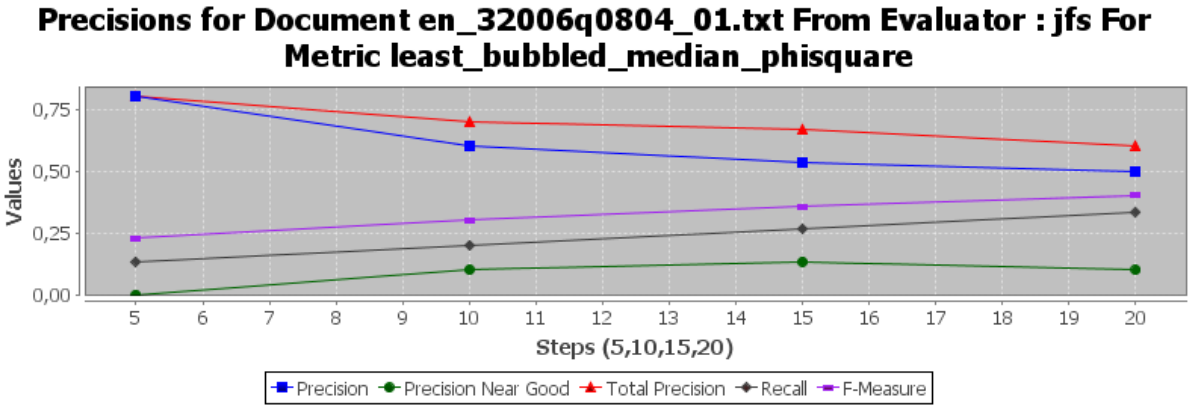


Figura 8.61 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Phi-Square

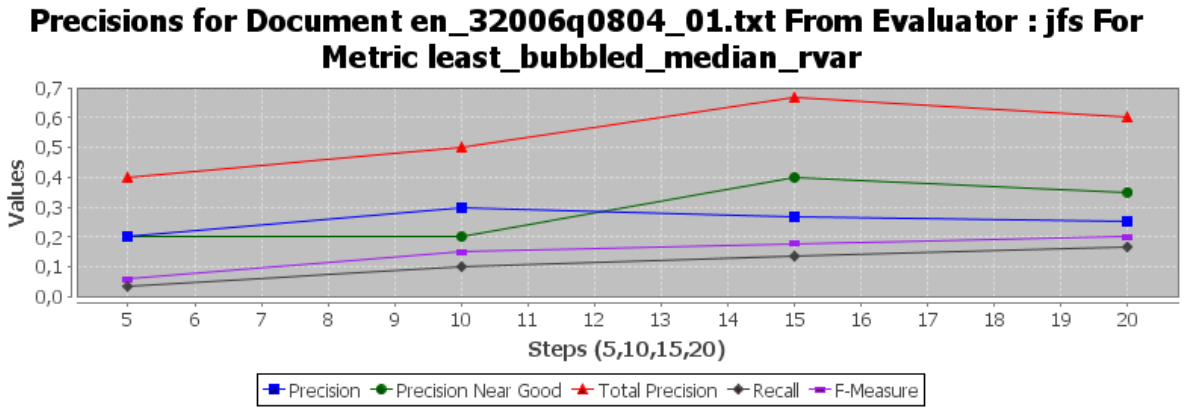


Figura 8.62 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Rvar

8.25 Gráficos da Precisão Total para todos os documentos em inglês avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva

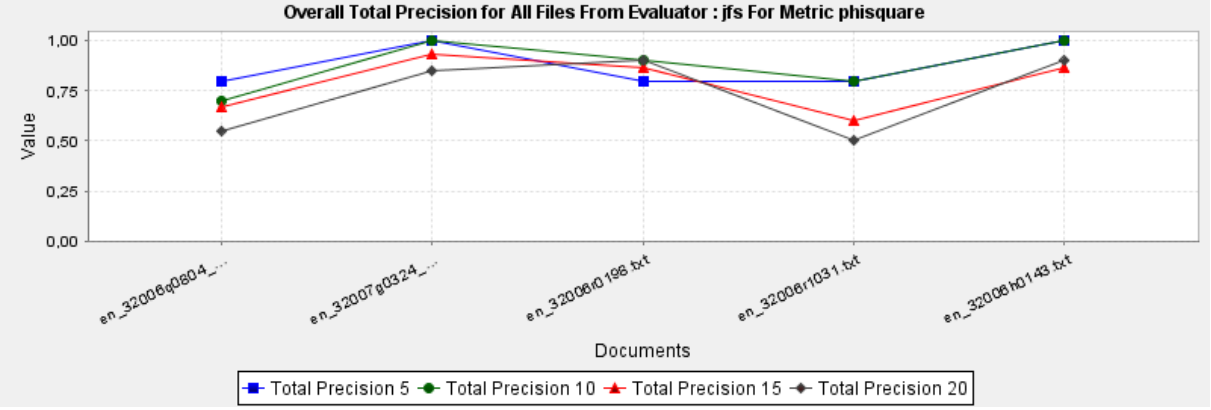


Figura 8.63 - Precisão total para todos os documentos em Inglês, para a medida Phi-Square

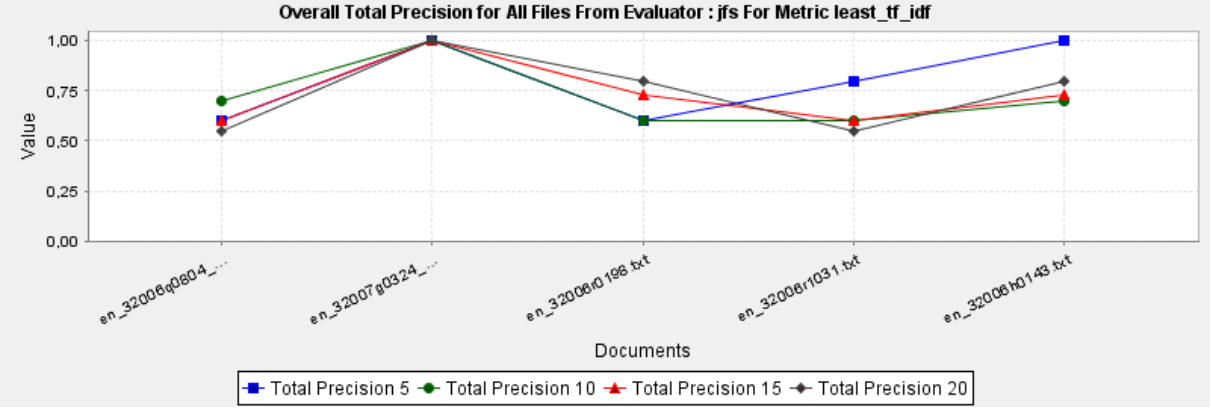


Figura 8.64 - Precisão total para todos os documentos em Inglês, para a medida Least Tf-Idf

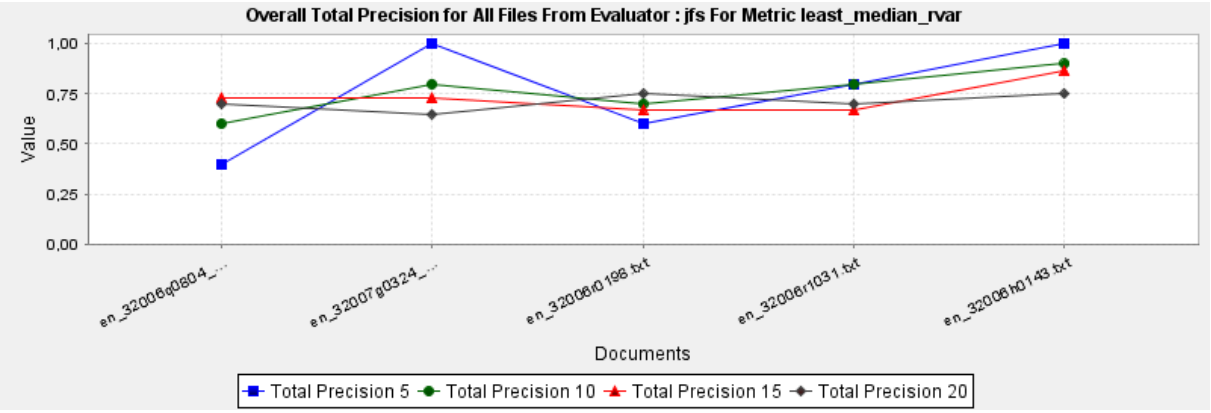


Figura 8.65 - Precisão total para todos os documentos em Inglês, para a medida Least Median Rvar

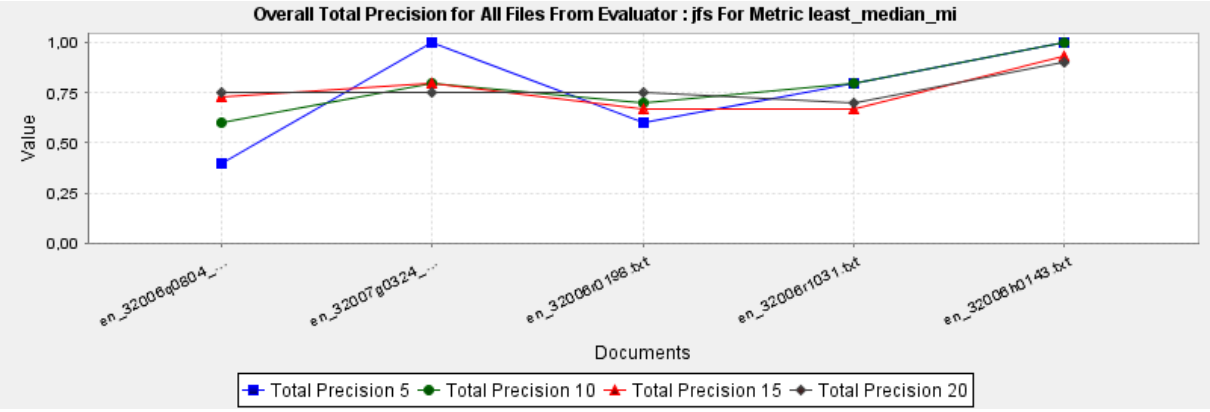


Figura 8.66 - Precisão total para todos os documentos em Inglês, para a medida Least Median MI

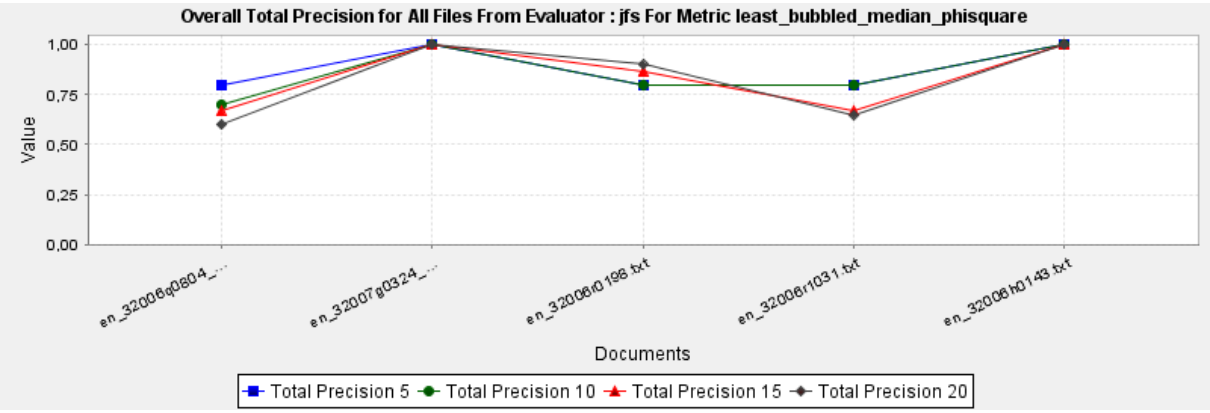


Figura 8.67 - Precisão total para todos os documentos em Inglês, para a medida Least Bubbled Median Phi-Square

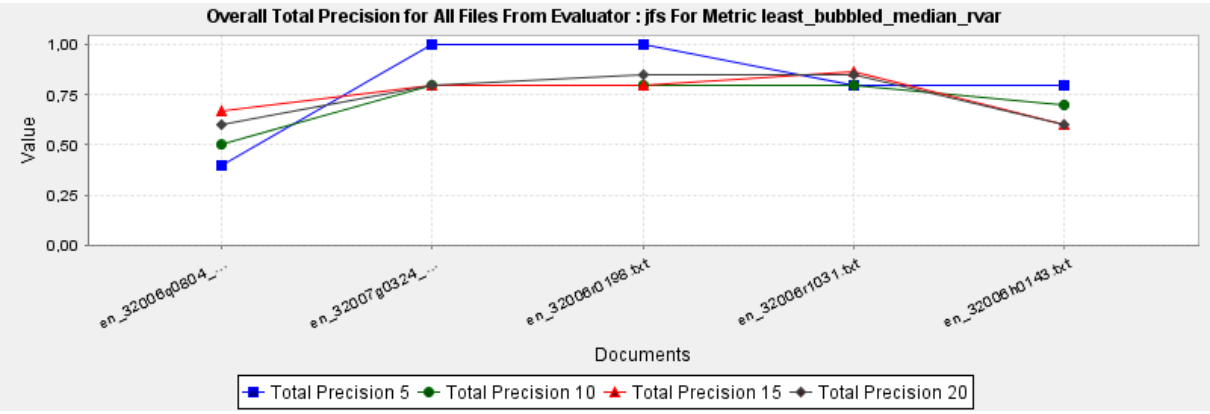


Figura 8.68 - Precisão total para todos os documentos em Inglês, para a medida Least Bubbled Median Rvar

8.26 Gráficos da Precisão Total versus Média da Precisão Total para todos os documentos em inglês avaliados pelo Avaliador Prof. Joaquim Ferreira da Silva

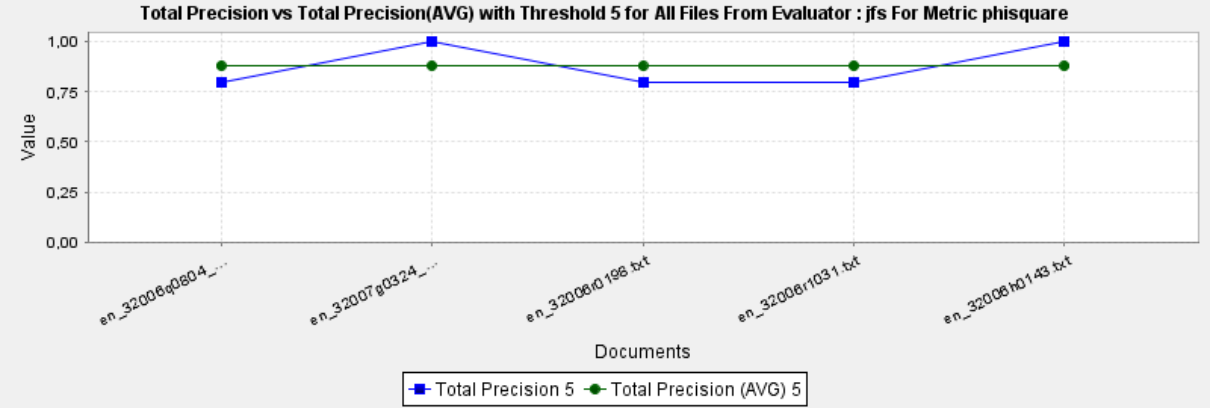


Figura 8.69 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 5

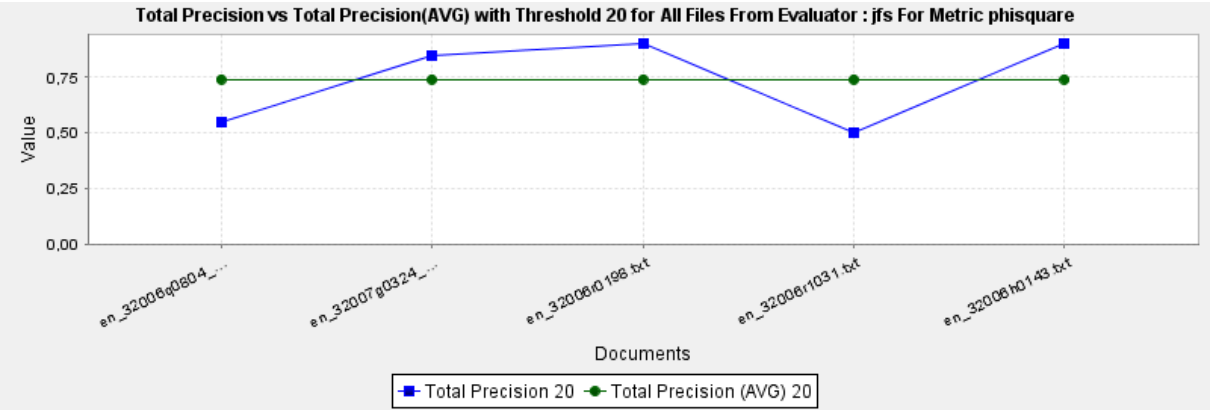


Figura 8.70 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Phi-Square, com o limite 20

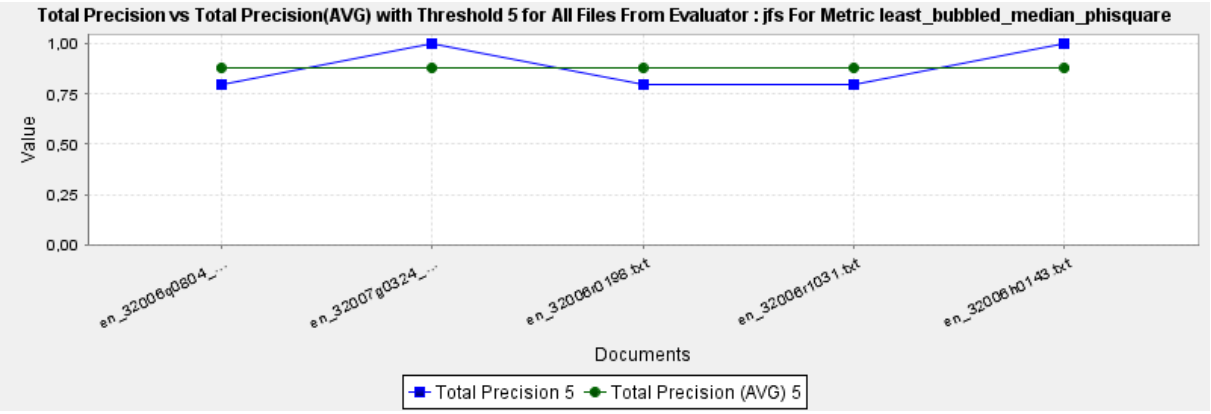


Figura 8.71 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 5

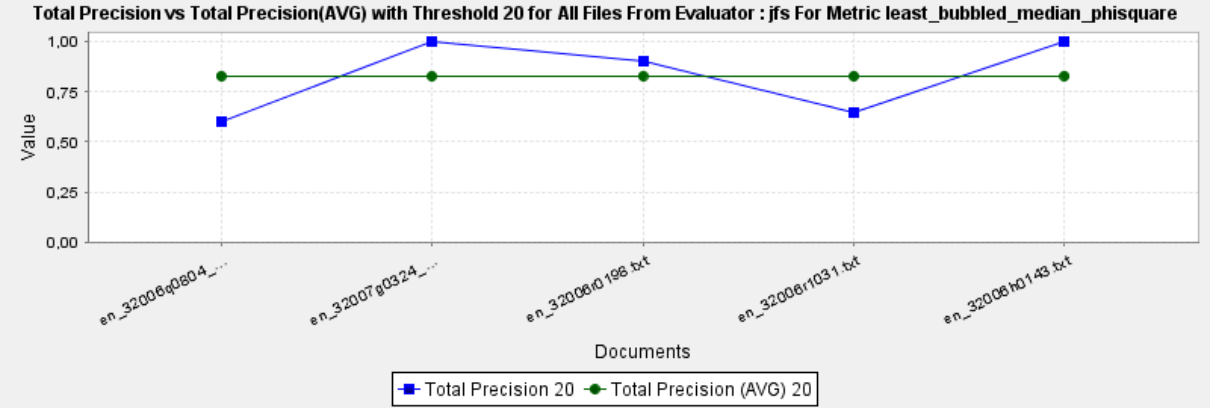


Figura 8.72 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Bubbled Median Phi-Square, com o limite 20

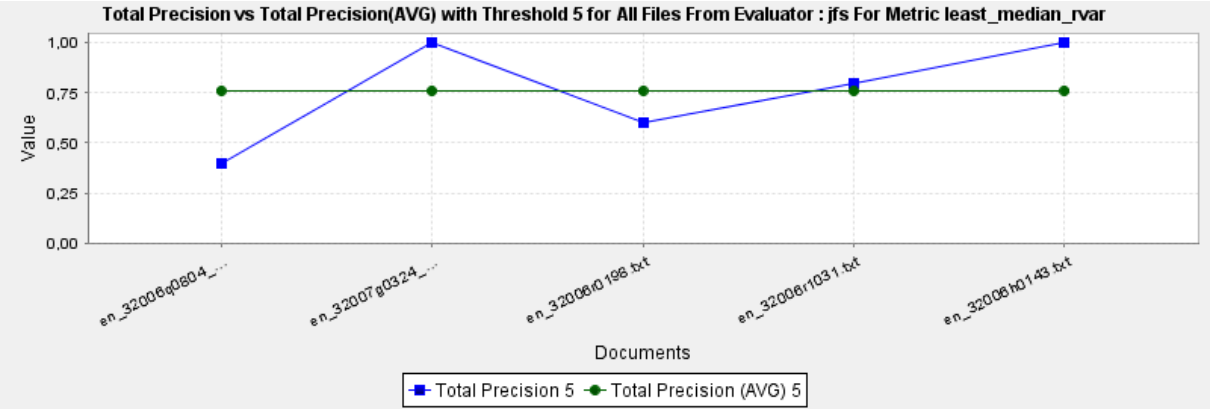


Figura 8.73 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 5

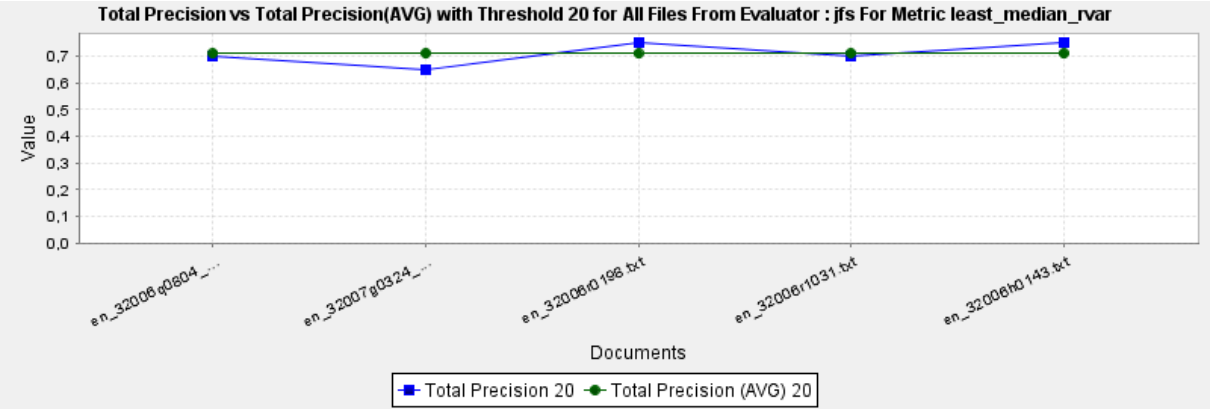


Figura 8.74 - Precisão total vs Precisão Total Média para todos os documentos, para a medida Least Median Rvar, com o limite 20

8.27 Tabela da Precisão Total Média para todas as Medidas resultante da Avaliação dos documentos em inglês pelo Avaliador Prof. Joaquim Ferreira da Silva

Metric	Precision Avg (5)	Precision Avg (10)	Precision Avg (15)	Precision Avg (20)
least_bubbled_median_rvar	0,8	0,72	0,746666667	0,74
least_bubbled_phisquare	0,84	0,78	0,783809524	0,738421053
phisquare	0,88	0,88	0,786666667	0,74
least_median_tf_idf	0,8	0,84	0,773333333	0,766315789
bubbled_phisquare	0,8	0,8	0,80952381	0,771351909
least_median_rvar	0,76	0,76	0,733333333	0,71
least_bubbled_median_mi	0,84	0,78	0,766666667	0,806986584
least_bubbled_tf_idf	0,88	0,78	0,747179487	0,738070175
least_median_phisquare	0,88	0,84	0,84	0,814736842
bubbled_mi	0,57	0,632619048	0,655104895	0,702352941
least_phisquare	0,84	0,837777778	0,811428571	0,74005848
least_median_mi	0,76	0,78	0,76	0,77
bubbled_rvar	0,57	0,605952381	0,645104895	0,685686275
least_tf_idf	0,8	0,72	0,733333333	0,74
least_bubbled_median_phisquare	0,88	0,86	0,84	0,83
least_bubbled_median_tf_idf	0,92	0,86	0,811428571	0,789649123
rvar	N/A	N/A	N/A	N/A
least_rvar	0,616666667	0,589285714	0,603982684	0,625
tf_idf	0,84	0,82	0,76	0,74
least_bubbled_rvar	N/A	0,605952381	0,654195804	0,685396825
mi	N/A	N/A	N/A	N/A
least_bubbled_mi	N/A	0,622619048	0,655104895	0,700784314
least_mi	0,55	0,564285714	0,583982684	0,625634921
bubbled_tf_idf	0,84	0,82	0,778754579	0,77166937

Tabela 8.61 - Precisão total média, para todas as medidas, resultante da avaliação do Avaliador Prof. Joaquim Ferreira da Silva

8.28 Tabela da Cobertura Média para todas as Medidas resultante da Avaliação dos documentos em inglês pelo Avaliador Prof. Joaquim Ferreira da Silva

Metric	Recall Avg (5)	Recall Avg (10)	Recall Avg (15)	Recall Avg (20)
least_bubbled_median_rvar	0,094288932	0,153563385	0,210683719	0,271460746
least_bubbled_phisquare	0,087804083	0,176313295	0,221560547	0,264425501
phisquare	0,11032156	0,188927434	0,232007919	0,291060625
least_median_tf_idf	0,090047673	0,173369155	0,211533075	0,261124219
bubbled_phisquare	0,082199688	0,145124165	0,184399375	0,232313348
least_median_rvar	0,074420563	0,143388847	0,204761635	0,240959922
least_bubbled_median_mi	0,070490465	0,128624219	0,194740627	0,284397759
least_bubbled_tf_idf	0,095224898	0,158378313	0,2167103	0,249040886
least_median_phisquare	0,102596693	0,177743213	0,248654385	0,29233301
bubbled_mi	0,036427225	0,048559578	0,069988149	0,109371903
least_phisquare	0,101058231	0,161712993	0,230192846	0,284019339
least_median_mi	0,068768095	0,131018627	0,199411562	0,24732508
bubbled_rvar	0,036427225	0,048559578	0,069988149	0,102705236
least_tf_idf	0,085379238	0,136075757	0,211789643	0,271135828
least_bubbled_median_phisquare	0,095255064	0,182920437	0,252938752	0,311211215
least_bubbled_median_tf_idf	0,083876858	0,163506518	0,215528981	0,267026234
rvar	0,013392857	0,013392857	0,020535714	0,020535714
least_rvar	0,017628205	0,04780543	0,053687783	0,079768369
tf_idf	0,110523863	0,178914027	0,230103157	0,280478615
least_bubbled_rvar	0,030544872	0,048559578	0,069988149	0,101474898
mi	0,013392857	0,013392857	0,020535714	0,020535714
least_bubbled_mi	0,030544872	0,042677225	0,069988149	0,110156216
least_mi	0,0125	0,04780543	0,053687783	0,073101702
bubbled_tf_idf	0,076703835	0,145878313	0,171066042	0,226430995

Tabela 8.62 - Cobertura média, para todas as medidas, resultante da avaliação do Avaliador Prof. Joaquim Ferreira da Silva

8.29 Lista de Termos Avaliados pelo Avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html

8.29.1 Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnost	0,007099977155724	good topic descriptor
podskupiny	0,005071328357677	near good descriptor
mnohojazyčnosti	0,005071328357677	good topic descriptor
skupiny	0,004070066317448	near good descriptor
vysoké úrovni pro mnohojazyčnost	0,004057029128410	good topic descriptor
skupina	0,003340670032842	near good descriptor
oblasti mnohojazyčnosti	0,003042746678425	good topic descriptor
pozorovatelům	0,002028481007305	near good descriptor
odbornou způsobilostí	0,002028481007305	good topic descriptor
zřízení skupiny na vysoké	0,002028481007305	near good descriptor
konzultovat	0,002028481007305	bad descriptor
skupinu konzultovat	0,002028481007305	bad descriptor
výdaje na zasedání	**0,002028481007305**	**unkonwn**
jména členů	0,002028481007305	bad descriptor
skupiny nebo podskupiny	0,002028481007305	near good descriptor
odborníkům a pozorovatelům	0,002028481007305	near good descriptor
skupina na vysoké	0,002028481007305	bad descriptor
osm až dvanáct	0,002028481007305	bad descriptor
skupině	0,002028481007305	near good descriptor
způsobilostí	0,002028481007305	bad descriptor
nahrazeni	0,002028481007305	bad descriptor
útvary	0,001341325852520	bad descriptor
skupiny na vysoké	0,001341325852520	bad descriptor
útvary komise	0,001341325852520	bad descriptor
odborníkům	0,001275871712364	near good descriptor

Tabela 8.63 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Phi-Square

8.29.2 Least Tf-Idf

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnost	0,025845015734672	good topic descriptor
podskupiny	0,018460725524766	near good descriptor
mnohojazyčnosti	0,018460725524766	good topic descriptor
skupina	0,013619695407680	near good descriptor
mnohojazyčnost zřizuje se skupina	0,013619695407680	bad descriptor
skupina a její podskupiny	0,013619695407680	near good descriptor
skupiny nebo podskupiny	0,012000622357528	near good descriptor
skupiny	0,012000622357528	near good descriptor
pozorovatelům	0,007384290209906	near good descriptor
konzultovat	0,007384290209906	bad descriptor
způsobilostí v oblasti mnohojazyčnosti	0,007384290209906	good topic descriptor
skupině	0,007384290209906	near good descriptor
způsobilostí	0,007384290209906	bad descriptor
nahrazeni	0,007384290209906	bad descriptor
odborníkům	0,006823998624308	near good descriptor
odborníkům a pozorovatelům	0,006823998624308	near good descriptor
skupina na vysoké	0,006263707038709	bad descriptor
vysoké úrovni pro mnohojazyčnost	0,006263707038709	good topic descriptor
skupiny na vysoké	0,006263707038709	bad descriptor
vysoké	0,006263707038709	bad descriptor
útvary	0,005966811313056	bad descriptor
skupině přidělily příslušné útvary	0,005966811313056	bad descriptor
funkčního období nahrazeni	0,005966811313056	bad descriptor
funkčního	0,005966811313056	bad descriptor
osobně	0,005966811313056	bad descriptor

Tabela 8.64 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Least Tf-Idf

8.29.3 Least Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnosti	15,000000000000000	good topic descriptor
mnohojazyčnost	14,000000000000000	good topic descriptor
projednávaných	14,000000000000000	bad descriptor
pozorovatelům	13,000000000000000	near good descriptor
způsobilostí	12,000000000000000	bad descriptor
zabezpečuje	11,000000000000000	bad descriptor
konzultovat	11,000000000000000	bad descriptor
shromažďují	11,000000000000000	bad descriptor
zabezpečuje sekretářské	11,000000000000000	bad descriptor
sekretářské	11,000000000000000	bad descriptor
pozorovatelům cestovní	10,500000000000000	bad descriptor
podskupiny	10,000000000000000	near good descriptor
prostorách	10,000000000000000	bad descriptor
nepřísluší	10,000000000000000	bad descriptor
neexistuje	10,000000000000000	bad descriptor
zveřejněna	10,000000000000000	bad descriptor
podskupiny budou rozpuštěny	10,000000000000000	bad descriptor
zveřejňují	10,000000000000000	bad descriptor
jednotlivě	10,000000000000000	bad descriptor
rozpuštěny	10,000000000000000	bad descriptor
způsobilostí v oblasti mnohojazyčnosti	9,500000000000000	good topic descriptor
důvěrných	9,000000000000000	bad descriptor
zveřejnit	9,000000000000000	bad descriptor
zůstávají	9,000000000000000	bad descriptor
původním jazyce dotyčného dokumentu zveřejnit	9,000000000000000	bad descriptor

Tabela 8.65 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Least Median Rvar

8.29.4 Least Median MI

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnosti	72,258388635426410	good topic descriptor
mnohojazyčnost	67,441162726397980	good topic descriptor
projednávaných	67,441162726397980	bad descriptor
pozorovatelům	62,623936817369554	near good descriptor
způsobilostí	57,806710908341130	bad descriptor
zabezpečuje	52,989484999312700	bad descriptor
konzultovat	52,989484999312700	bad descriptor
shromažďují	52,989484999312700	bad descriptor
zabezpečuje sekretářské	52,989484999312700	bad descriptor
sekretářské	52,989484999312700	bad descriptor
pozorovatelům cestovní	50,580872044798490	bad descriptor
pozorovatele	49,488944741621780	near good descriptor
zpracovávají	49,488944741621780	bad descriptor
podskupiny	48,172259090284270	bad descriptor
prostorách	48,172259090284270	bad descriptor
nepřísluší	48,172259090284270	bad descriptor
neexistuje	48,172259090284270	bad descriptor
zveřejněna	48,172259090284270	bad descriptor
podskupiny budou rozpuštěny	48,172259090284270	bad descriptor
zveřejňují	48,172259090284270	bad descriptor
jednotlivě	48,172259090284270	good topic descriptor
rozpuštěny	48,172259090284270	bad descriptor
způsobilostí v oblasti mnohojazyčnosti	45,763646135770060	bad descriptor
zveřejňování	44,623363444323815	bad descriptor
spravováno_úřadem_pro_úřední_tisky	43,889423070017045	bad descriptor

Tabela 8.66 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Least Median MI

8.29.5 Least Bubbled Median Phi-Square

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnosti	0,168306300320869	good topic descriptor
mnohojazyčnost	0,157085880299478	good topic descriptor
podskupiny	0,060856443666432	near good descriptor
podskupin	0,054770799299789	near good descriptor
skupinou	0,051735406981616	near good descriptor
skupina	0,045268481108914	near good descriptor
skupinu	0,045268481108914	near good descriptor
mnohojazyčnost zřizuje se skupina	0,045268481108914	bad descriptor
skupině	0,045268481108914	near good descriptor
skupiny	0,045268481108914	near good descriptor
skupiny nebo podskupiny	0,042599510566502	near good descriptor
skupin	0,038801555236212	near good descriptor
skupin a podskupin	0,036513866199859	near good descriptor
skupina a její podskupiny	0,033471044016537	near good descriptor
zveřejňování	0,011201716547091	bad descriptor
skupině přidělily příslušné útvary	0,010730606820159	bad descriptor
pozorovatelům	0,010324493133595	near good descriptor
nepřísluší	0,010142321146361	bad descriptor
neexistuje	0,010142321146361	bad descriptor
podskupiny budou rozpuštěny	0,010142321146361	bad descriptor
rozpuštěny	0,010142321146361	bad descriptor
pozorovatele	0,009530301354087	near good descriptor
zveřejnění	0,009334763789243	bad descriptor
zveřejněna	0,009334763789243	bad descriptor
zveřejňují	0,009334763789243	bad descriptor

Tabela 8.67 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Least Bubbled Median Phi-Square

8.29.6 Least Bubbled Median Rvar

Termos	Valor da Medida	Avaliação dada ao termo pelo Avaliador
mnohojazyčnosti	14,570893949858611	good topic descriptor
mnohojazyčnost	13,599501019868036	good topic descriptor
podskupiny	10,000000000000000	near good descriptor
nepřísluší	10,000000000000000	bad descriptor
neexistuje	10,000000000000000	bad descriptor
podskupiny budou rozpuštěny	10,000000000000000	bad descriptor
rozpuštěny	10,000000000000000	bad descriptor
vyzrazeny	9,000000000000000	bad descriptor
podskupin	9,000000000000000	near good descriptor
podskupiny nesmějí být vyzrazeny	8,000000000000000	bad descriptor
nepřísluší odměna	8,000000000000000	bad descriptor
nedodrží	8,000000000000000	bad descriptor
pozorovatelům	7,709636786377628	near good descriptor
zabezpečuje	7,315962630517282	bad descriptor
pozorovatele	7,116587802810118	near good descriptor
vlivech	7,000000000000000	bad descriptor
tématem	7,000000000000000	bad descriptor
dodávat	7,000000000000000	bad descriptor
dodávat nové podněty a nápady	6,000000000000000	bad descriptor
nápady	6,000000000000000	bad descriptor
usoudí	6,000000000000000	bad descriptor
uhradí	6,000000000000000	bad descriptor
limitů	6,000000000000000	bad descriptor
zřídit	6,000000000000000	bad descriptor
odměna	6,000000000000000	bad descriptor

Tabela 8.68 - Listagem de termos com as respectivas avaliações feitas pelo avaliador Prof. Gabriel Lopes para o documento cs_32006D0644.html na medida Least Bubbled Median Rvar

8.30 Lista de Termos Apresentados aos Avaliadores para outras medidas

8.30.1 Rvar

Termos	Valor da Medida
pokud komise usoudí	1,00
zřízeny podskupiny	1,00
kteří mají zájem na projednávaných	1,00
vhodné v určité otázce	1,00
mnohojazyčnosti v evropské unii	1,00
také prohlášení	1,00
nesmějí být vyzrazeny	1,00
pravidla o zveřejňování	1,00
podskupiny obvykle zasedají	1,00
zbývající část svého funkčního	1,00
mnohojazyčnost v souladu se sdělením	1,00
rozvrhem stanovenými komisí	1,00
mezích limitů ročního	1,00
učinit z pravidla o zveřejňování	1,00
souladu se sdělením komise nazvaným	1,00
mohl ohrozit jejich nezávislost	1,00
zavazují jednat	1,00
souvislosti s určitým tématem	1,00
odpovědný za mnohojazyčnost	1,00
zástupce komise požádat odborníky	1,00
úkoly úkolem skupiny	1,00
vlivech	1,00
pomáhat poskytovat podporu a poradenství	1,00
zabezpečuje	1,00
zda existuje či neexistuje zájem	1,00

Tabela 8.69 - Lista de Termos para a medida Rvar para o ficheiro cs_32006D0644.html

8.30.2 MI

Temos	Valor da Medida
pokud komise usoudí	4,8172259
zřízeny podskupiny	4,8172259
kteří mají zájem na projednávaných	4,8172259
vhodné v určité otázce	4,8172259
mnohojazyčnosti v evropské unii	4,8172259
také prohlášení	4,8172259
nesmějí být vyzrazeny	4,8172259
pravidla o zveřejňování	4,8172259
podskupiny obvykle zasedají	4,8172259
zbývající část svého funkčního	4,8172259
mnohojazyčnost v souladu se sdělením	4,8172259
rozvrhem stanovenými komisí	4,8172259
mezích limitů ročního	4,8172259
učinit z pravidla o zveřejňování	4,8172259
souladu se sdělením komise nazvaným	4,8172259
mohl ohrozit jejich nezávislost	4,8172259
zavazují jednat	4,8172259
souvislosti s určitým tématem	4,8172259
odpovědný za mnohojazyčnost	4,8172259
zástupce komise požádat odborníky	4,8172259
úkoly úkolem skupiny	4,8172259
vlivech	4,8172259
pomáhat poskytovat podporu a poradenství	4,8172259
zabezpečuje	4,8172259
zda existuje či neexistuje zájem	4,8172259

Tabela 8.70 - Lista de Termos para a medida MI para o ficheiro cs_32006D0644.html

8.30.3 Tf-Idf

Termo	Valor da Medida
mnohojazyčnost	0,025845
podskupiny	0,0184607
mnohojazyčnosti	0,0184607
vysoké úrovni pro mnohojazyčnost	0,0147686
skupina	0,0136197
skupiny	0,0120006
oblasti mnohojazyčnosti	0,0110764
pozorovatelům	0,0073843
odbornou způsobilostí	0,0073843
zřízení skupiny na vysoké	0,0073843
konzultovat	0,0073843
skupinu konzultovat	0,0073843
výdaje na zasedání	0,0073843
jména členů	0,0073843
skupiny nebo podskupiny	0,0073843
odborníkům a pozorovatelům	0,0073843
skupina na vysoké	0,0073843
osm až dvanáct	0,0073843
skupině	0,0073843
způsobilostí	0,0073843
nahrazeni	0,0073843
odborníkům	0,006824
vysoké	0,0062637
útvary	0,0059668
skupiny na vysoké	0,0059668

Tabela 8.71 - Lista de Termos para a medida Tf-Idf para o ficheiro cs_32006D0644.html

8.31 Gráficos das Precisões para o Prof. Gabriel Lopes para o documento cs_32006D0644.html

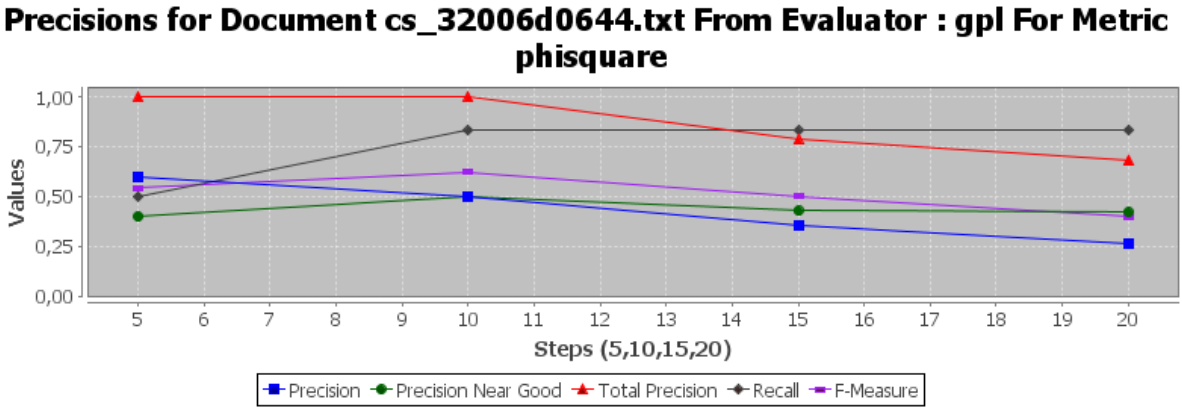


Figura 8.75 - Valores de Precisão, Cobertura e F-Measure para Phi-Square

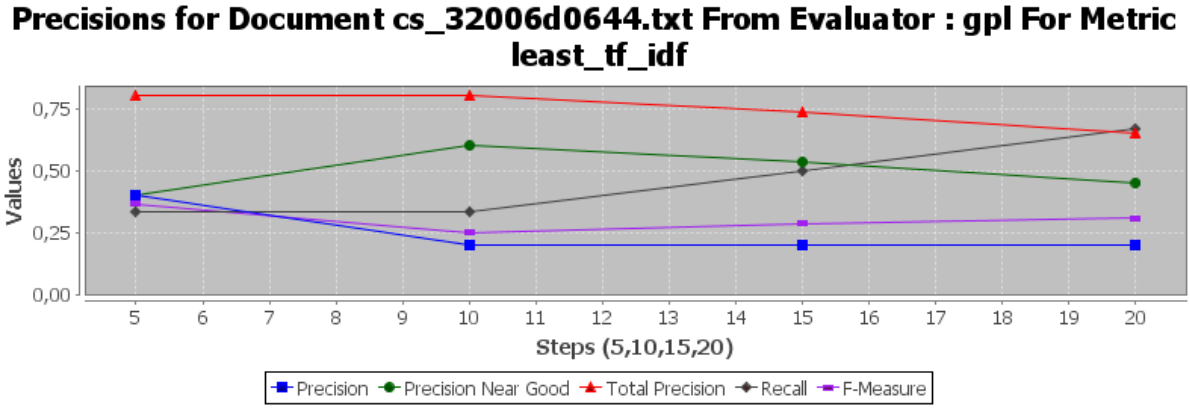


Figura 8.76 - Valores de Precisão, Cobertura e F-Measure para Least Tf-Idf

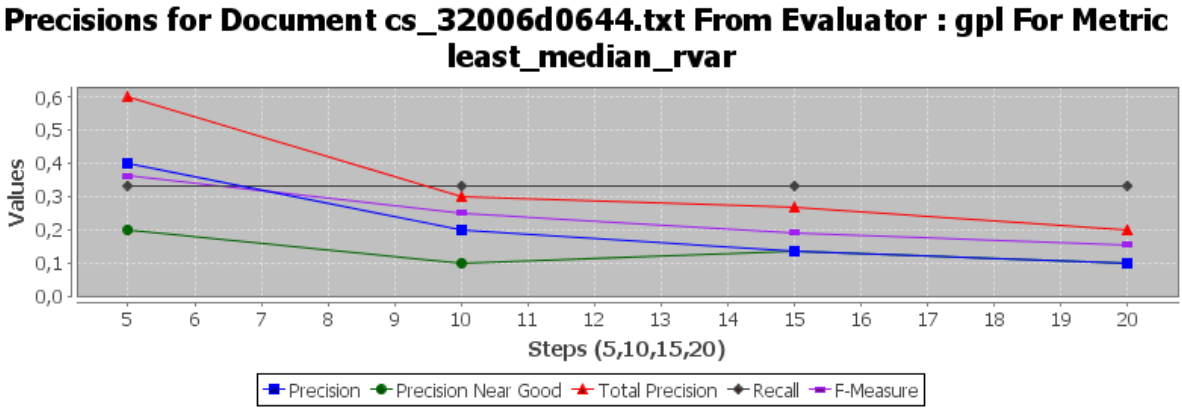


Figura 8.77 - Valores de Precisão, Cobertura e F-Measure para Least Median Rvar

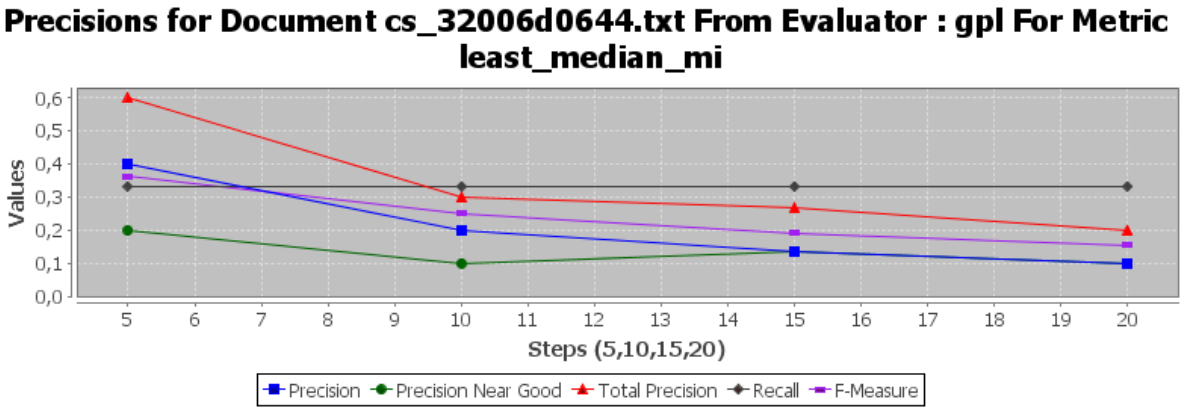


Figura 8.78 - Valores de Precisão, Cobertura e F-Measure para Least Median MI

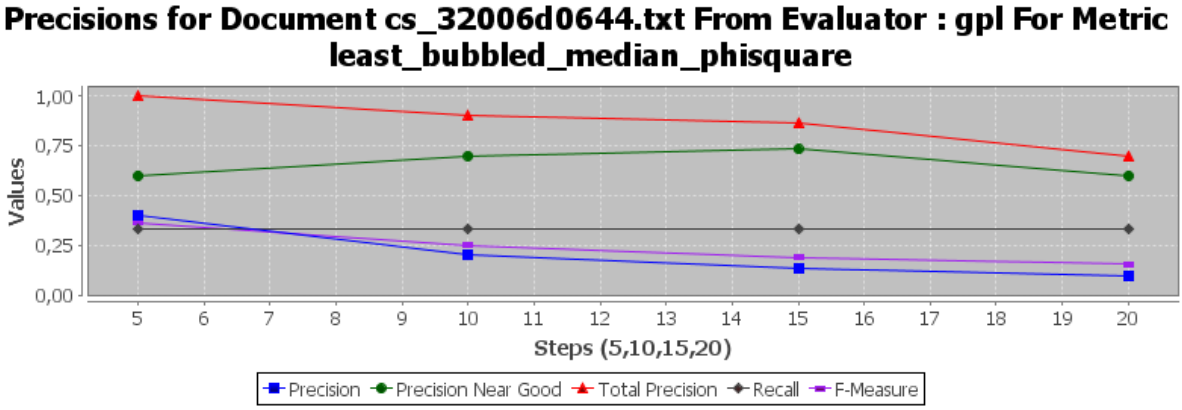


Figura 8.79 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Phi-Square

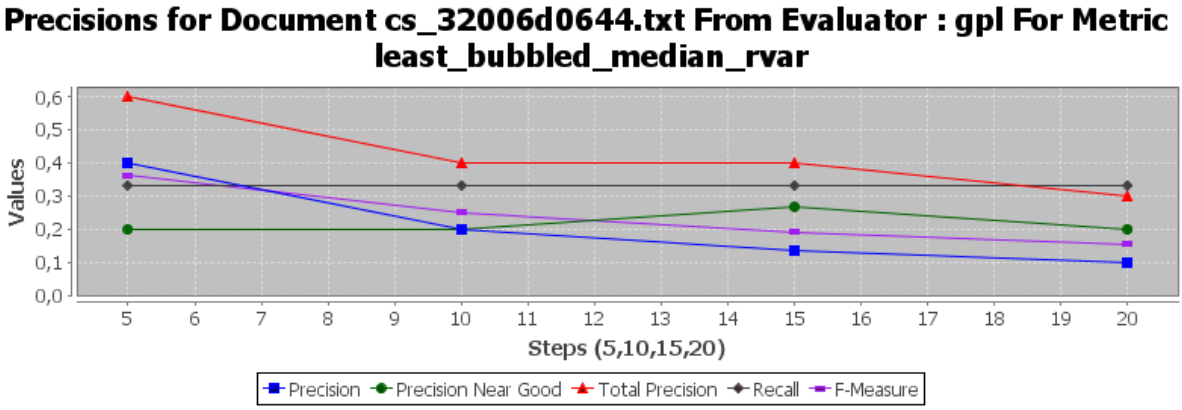


Figura 8.80 - Valores de Precisão, Cobertura e F-Measure para Least Bubbled Median Rvar

8.32 Gráficos da Precisão Total para todos os documentos em Checo avaliados pelo Avaliador Prof. Gabriel Lopes

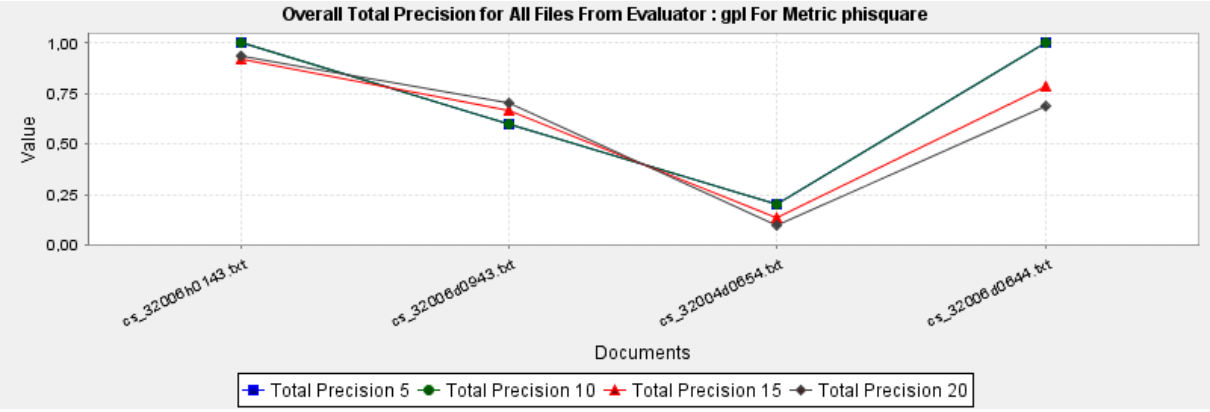


Figura 8.81 - Precisão total para todos os documentos em Checo, para a medida Phi-Square

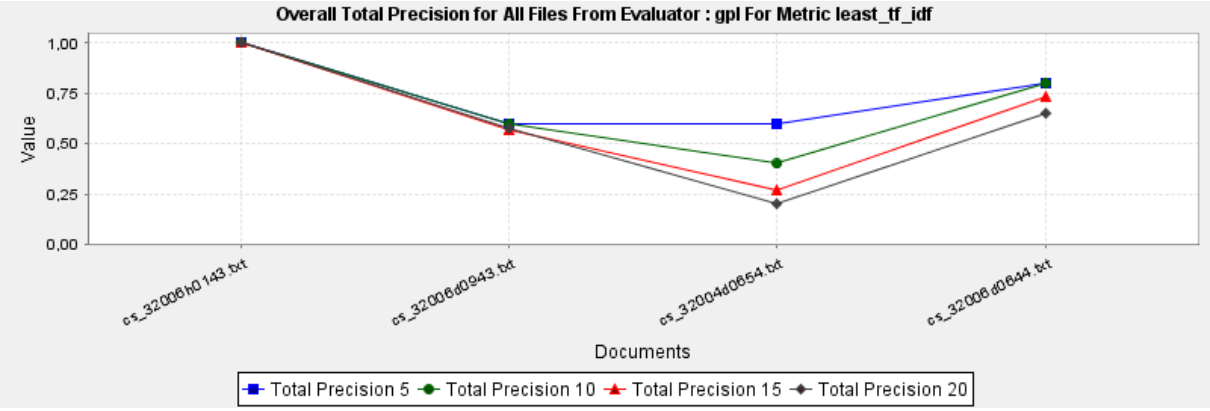


Figura 8.82 - Precisão total para todos os documentos em Checo, para a medida Least Tf-Idf

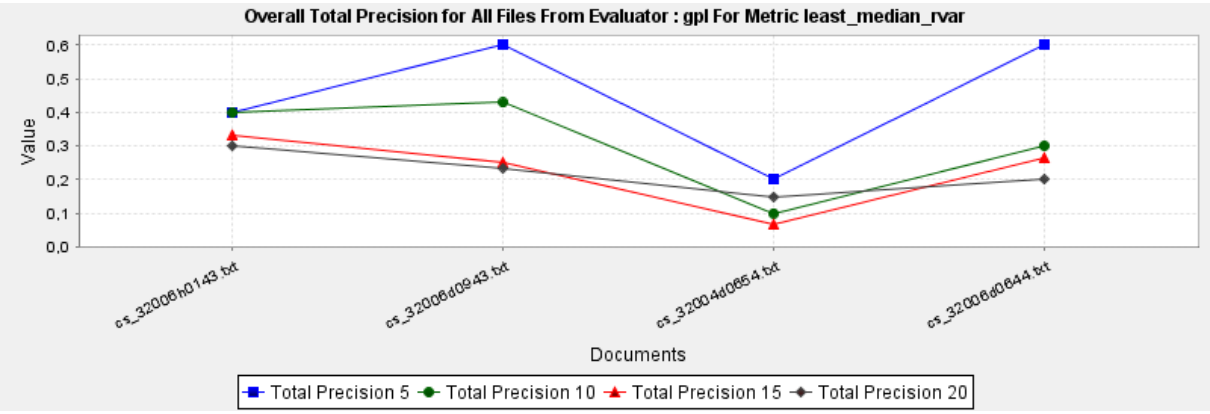


Figura 8.83 - Precisão total para todos os documentos em Checo, para a medida Least Median Rvar

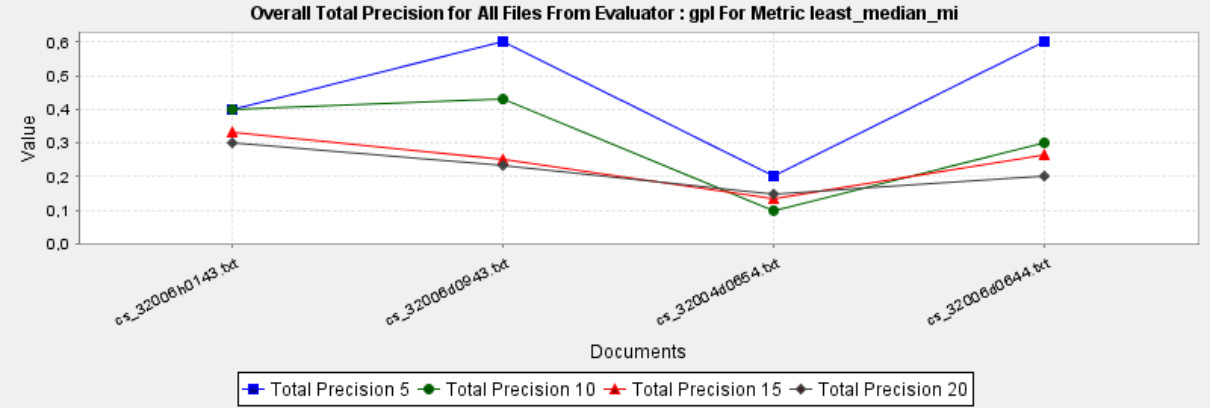


Figura 8.84 - Precisão total para todos os documentos em Checo, para a medida Least Median MI

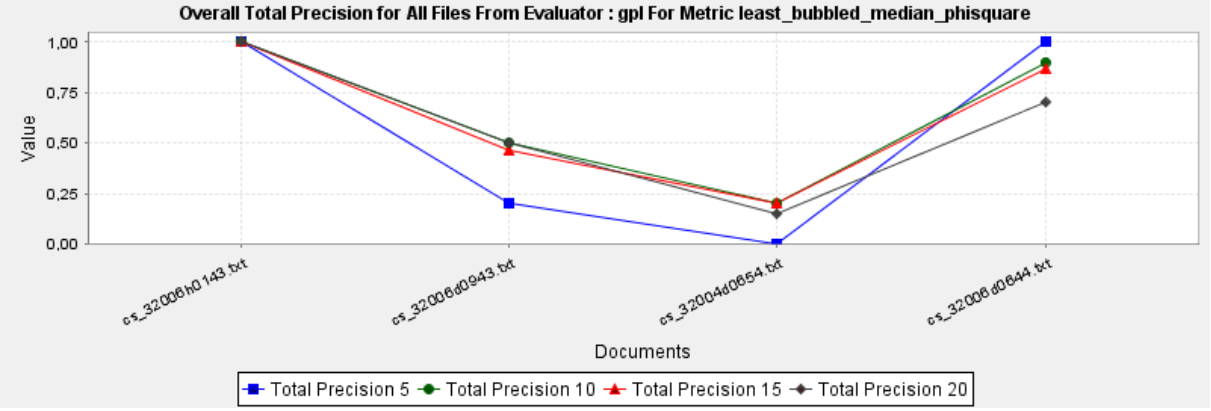


Figura 8.85 - Precisão total para todos os documentos em Checo, para a medida Least Bubbled Median Phi-Square

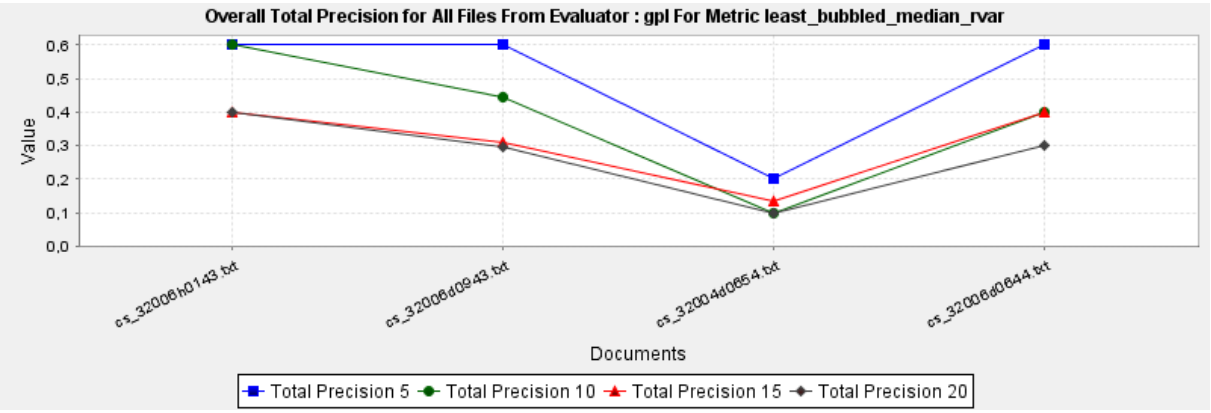


Figura 8.86 - Precisão total para todos os documentos em Checo, para a medida Least Bubbled Median Rvar

8.33 Tabela da Precisão Total Média para todas as Medidas resultante da Avaliação dos documentos em Checo pelo Avaliador Prof. Gabriel Lopes

Metric	Prec. Avg (5)	Prec. Avg (10)	Prec. Avg (15)	Prec. Avg (20)
least_bubbled_median_rvar	0,5	0,386111111	0,31025641	0,273529412
least_bubbled_phisquare	0,55	0,625	0,633333333	0,555263158
phisquare	0,7	0,7	0,625595238	0,605427632
least_median_tf_idf	0,7	0,65	0,633333333	0,5875
bubbled_phisquare	0,6	0,675	0,566666667	0,5125
least_median_rvar	0,45	0,307142857	0,229166667	0,221323529
least_bubbled_median_mi	0,4	0,4	0,312820513	0,263596491
least_bubbled_tf_idf	0,8	0,675	0,666666667	0,651388889
least_median_phisquare	0,7	0,6	0,566666667	0,575
bubbled_mi	0,175	0,2	0,204778555	0,19540036
least_phisquare	0,7	0,6	0,566666667	0,582894737
least_median_mi	0,45	0,307142857	0,245833333	0,221323529
bubbled_rvar	0,225	0,2	0,151893939	0,202727501
least_tf_idf	0,75	0,7	0,642857143	0,607236842
least_bubbled_median_phisquare	0,55	0,65	0,633333333	0,5875
least_bubbled_median_tf_idf	0,65	0,675	0,7	0,664667183
rvar	N/A	N/A	N/A	N/A
least_rvar	0,258333333	0,251984127	0,227083333	0,214239927
tf_idf	0,9	0,85625	0,709249084	0,659813596
least_bubbled_rvar	0,175	0,20625	0,156060606	0,169966063
mi	N/A	N/A	N/A	N/A
least_bubbled_mi	0,125	0,20625	0,168881119	0,206730769
least_mi	0,258333333	0,251984127	0,227083333	0,214239927
bubbled_tf_idf	0,75	0,675	0,625	0,598611111

Tabela 8.72- Precisão total média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

8.34 Tabela da Cobertura Média para todas as Medidas resultante da Avaliação dos documentos em Checo pelo Avaliador Prof. Gabriel Lopes

Metric	Recall Avg (5)	Recall Avg (10)	Recall Avg (15)	Recall Avg (20)
least_bubbled_median_rvar	0,172443978	0,184348739	0,184348739	0,196253501
least_bubbled_phisquare	0,083333333	0,18697479	0,37710084	0,400910364
phisquare	0,192927171	0,391981793	0,421393557	0,504026611
least_median_tf_idf	0,18714986	0,362570028	0,472163866	0,525385154
bubbled_phisquare	0,083333333	0,213760504	0,228466387	0,243172269
least_median_rvar	0,160539216	0,160539216	0,172443978	0,246848739
least_bubbled_median_mi	0,098039216	0,184348739	0,184348739	0,196253501
least_bubbled_tf_idf	0,175245098	0,31197479	0,454306723	0,596988796
least_median_phisquare	0,199054622	0,266981793	0,335259104	0,510679272
bubbled_mi	0,014705882	0,089110644	0,089110644	0,18714986
least_phisquare	0,210959384	0,25227591	0,394957983	0,49877451
least_median_mi	0,160539216	0,160539216	0,172443978	0,246848739
bubbled_rvar	0,077205882	0,089110644	0,089110644	0,18714986
least_tf_idf	0,261554622	0,380077031	0,474964986	0,569852941
least_bubbled_median_phisquare	0,083333333	0,24947479	0,389005602	0,427521008
least_bubbled_median_tf_idf	0,083333333	0,288165266	0,454481793	0,543417367
rvar	0,011904762	0,089110644	0,089110644	0,089110644
least_rvar	0,077205882	0,130777311	0,130777311	0,205182073
tf_idf	0,344537815	0,546393557	0,620798319	0,676820728
least_bubbled_rvar	0,077205882	0,089110644	0,089110644	0,151610644
mi	0,011904762	0,089110644	0,089110644	0,089110644
least_bubbled_mi	0,014705882	0,089110644	0,089110644	0,181022409
least_mi	0,077205882	0,130777311	0,130777311	0,205182073
bubbled_tf_idf	0,160539216	0,213760504	0,290966387	0,290966387

Tabela 8.73 - Cobertura média, para todas as medidas, resultante da avaliação do Avaliador Prof. Gabriel Lopes

Bibliografia

[1] J. F. d. Silva, and G. P. Lopes, “A Document Descriptor Extractor Based on Relevant Expressions,” in 14th Portuguese Conference on Artificial Intelligence, EPIA 2009, Aveiro, Portugal, October 12-15, 2009, pp. 646-657.

[2] J. F. d. Silva, G. Dias, S. Guilloré *et al.*, “Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units,” in 9th Portuguese conference on artificial intelligence Evora, 21-24 September 1999 1999.

[3] D. Franca, and S. Fabrizio, “Supervised term weighting for automated text categorization,” in Proceedings of the 2003 ACM symposium on Applied computing, Melbourne, Florida, 2003.

[4] Y. Yiming, and O. P. Jan, “A Comparative Study on Feature Selection in Text Categorization,” in Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

[5] F. A. P. Madureira, “Classificação de Documentos,” Departamento de Informática, Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa, Lisboa, 2009.

[6] F. Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[7] J. F. d. Silva, and G. P. Lopes, “Towards Automatic Building of Document Keywords,” in COLING 2010 - The 23rd International Conference on Computational Linguistics, Pequim, 2010.

[8] M. Yamamoto, and K. W. Church, "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus," pp. 1-30: Association for Computational Linguistics, 2001.

[9] R. Papka, and J. Allan, “Document classification using multiword features,” in Proceedings of the seventh international conference on Information and knowledge management, 1998, pp. 124-131.

[10] C. Jacquemin, *Spotting and discovering terms through natural language processing*: MIT Press, 2001.

[11] F. Geraci, M. Pellegrini, P. Pisati *et al.*, “A scalable algorithm for high-quality clustering of web snippets,” in Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, 2006.

[12] D. J. M. Ferreira, “Procura Estruturada de Textos para perfis de Utilizadores,” Departamento de Informática, Universidade da Beira Interior, 2009.

[13] P. Ferragina, and A. Gulli, “A personalized search engine based on web-snippet hierarchical clustering,” in Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan, 2005.

[14] F. Fukumoto, and Y. Suzuki, "Extracting Key Paragraph based on Topic and Event Detection -- Towards Multi-Document Summarization," *In Hahn et al*, pp. 31-39.
[15] J. F. d. Silva, and G. P. Lopes, "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units," in Proceedings of the 6th Meeting on the Mathematics of Language, Orlando, 1999, pp. 369-381.
[16] J. M. Cigarrán, A. Peñas, J. Gonzalo *et al.*, "Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retreival System," *ICFCA 2005*, B. Ganter and R. Godin, eds., p. 4963: Springer Berlin, 2005.
[17] J. M. Cigarrán, J. Gonzalo, A. Peñas *et al.*, "Browsing Search Results via Formal Concept Analysis: Automatic Selection of Attributes " *Concept Lattices*, Lecture Notes in Computer Science, pp. 201-202: Springer Berlin / Heidelberg, 2004.
[18] J. Hereth, G. Stumme, R. Wille *et al.*, "Conceptual knowledge discovery - a human-centered approach.," *Journal of Applied Artificial Intelligence*, vol. 17, no. 3, pp. 288–301, 2003.
[19] U. Priss, "Formal concept analysis in information science," *Information Science and Technology*, vol. 40, pp. 521–543, 2006.
[20] P. G. Otero, G. P. Lopes, and A. Agustini, "Automatic Acquisition of Formal Concepts from Text," *LDV-Forum*, vol. 23, no. 1, pp. 59-14, 2008.
[21] S. Gerard, and B. Chris, *Term Weighting Approaches in Automatic Text Retrieval*, Cornell University, 1987.
[22] G. Dias, "Extraction Automatique dSAssociations Lexicales à partir de Copora," Universidade Nova de Lisboa e LIFO Universidade de Orleans(França), Lisboa, Portugal, 2002.
[23] T. Afrin, "Extraction of Basic Noun Phrases from Natural Language Using Statistical Context-Free Grammar," Electrical Engineering, Virginia Polytechnic Institute and State University, 2001.
[24] J. L. Martínez-Fernández, A. García-Serrano, P. Martínez *et al.*, "Automatic Keyword Extraction for News Finder," *Adaptive Multimedia Retrieval*, Lecture Notes in Computer Science, pp. 405-427: Springer Berlin / Heidelberg, 2004.
[25] Y. Gao, and G. Zhao, "Knowledge-based Information Extraction: A case study of recognizing emails of Nigerian frauds," *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science: Springer Berlin / Heidelberg, 2005.
[26] A. Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge," pp. 216 - 223.
[27] J. M. J. Ventura, "Extracção de Unigramas Relevantes," Departamento de Informática, Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa, Lisboa, 2008.
[28] Y. Matsuo, and M. Ishizuka, "Keyword Extraction from a single Document using word Co-Occurence Statistical Information," *International Journal on Articial Intelligence Tools*, vol. 13, no. 1, pp. 157-169, 2004.
[29] K. Gurney, *An Introduction to Neural Networks*: CRC Press, 2003.
[30] A. Das, M. Marko, A. Probst *et al.*, "Neural Net Model for featured word extraction," *CoRR, cs. NE/0206001*, 2002.
[31] R. Yangarber, and R. Grishman, "Machine Learning of Extraction Patterns from Unannotated Corpora: Position Statement," *Workshop Machine Learning for Information Extraction*, I. Press, ed., pp. 76-83, Amsterdam, 2000.
[32] B. Georgantopoulos, and S. Piperidis, "Automatic Acquisition of Terminological Resources for Information Extraction Applications," in NIT Conference, Athens, 1998.
[33] A.-C. N. Ngomo, "Knowledge-Free Discovery of Domain-Specific Multiword Units," in SAC'08, Ceará, Brazil, 2008.

[34] Y. Uzun, "Keyword Extraction Using Naive Bayes," Bilkent University, Department of Computer Science, Turkey University, 2005.
[35] M. Litvak, and M. Last, "Graph-Based Keyword Extraction for Single-Document Summarization." pp. 17-24.
[36] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi, "Enhanced Web Document Summarization Using Hyperlinks." pp. 208 - 215.
[37] J. Allan, J. Carbonell, G. Doddington *et al.*, "Topic Detection and Tracking Pilot Study - Final Report."
[38] N. Guarino, "Formal Ontology and Information Systems," in Proceedings of FOIS'98, Trento, Italy, 1998, pp. 3-15.
[39] P. Velardi, M. MissiKoff, and R. Basili, "Identification of relevant Terms to support the construction of Domain Ontologies," in Proceedings of the workshop on Human Language Technology and Knowledge Management - Volume 2001, Tolouse, France, 2001.
[40] B. Fortuna, N. Lavrač, and P. Velardi, "Advancing Topic Ontology Learning Through Term Extraction," *PRICAI 2008: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, pp. 626-635: Springer Berlin / Heidelberg, 2008.
[41] B. Fortuna, M. Grobelnik, and D. Mladenič, "System for semi-automatic ontology construction," in 3rd Annual European Semantic Web Conference, Budva, Montenegro, 2006.
[42] B. Fortuna, D. Mladenic, and M. globelnik, "Semi-automatic Construction of Topic Ontologies," *Semantics, Web and Mining*, Lecture Notes in Computer Science, pp. 121-131: Springer Berlin / Heidelberg, 2006.
[43] J. Brank, D. Mladenić, M. Grobelnik *et al.*, "Feature selection using support vector machines," in Proceedings of the Third International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields, Bologna,Italy, 2002, pp. 25-27.
[44] T. Joachims, "Making large-scale svm learning practical," *Advances in Kernel Methods - Support Vector Learning.*, C. B. B. Scholkopf, and A. Smola, ed.: MIT-Press, 1999.
[45] A. Dingli, F. Ciravegna, D. Guthrie *et al.*, "Mining Web Sites Using Unsupervised Adaptive Information Extraction," in Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003.
[46] H. Alani, S. Kim, D. E. Millard *et al.*, "Automatic Extraction of Knowledge from Web Documents."
[47] U. Manber, and G. Myers, "Suffix arrays: A new method for on-line string searches," *SIAM Journal on Computing*, vol. 22, no. 5, pp. 935-948, 1993.
[48] K. Sadakane, "Compressed Suffix Trees with Full Functionality," in Theory Comput. Syst. 41(4), 2007, pp. 589-607.
[49] L. Russo, G. Navarro, and A. L. Oliveira, "Fully-Compressed Suffix Trees," *Lecture Notes on Computer Science* P. LATIN'08., ed., pp. 362-373, Berlin, Germany: Springer-Verlag, 2008.
[50] S. Burkhardt, and J. Karkkainen, "Fast Lightweight Suffix Array Construction and Checking," *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching*, pp. 55-69: Springer Berlin / Heidelberg, 2003.
[51] M. D. McIlroy. "Suffix arrays," http://www.cs.dartmouth.edu/~doug/sarray/.